基于加权合成少数类过采样技术的故障诊断

韩志艳,王 健
(渤海大学 工学院,辽宁 锦州 121000)

摘 要:合成少数类过采样技术(Synthetic Minority Oversampling Technique,SMOTE)是一种著名的过采样方法,但是它没有考虑样本的分布和潜在的噪声数据。为了改善 SMOTE 的性能,提出了加权合成少数类过采样技术(Weighted Synthetic Minority Oversampling Technique,WSMOTE)。WSMOTE 通过引入邻域并将样本按照分布的不同划分为不同的组群,不同的群组拥有不同的采样价值,然后根据采样价值的不同加权合成样本。WSMOTE 在处理类别不平衡数据时具有优异的性能,并在半导体制造过程的监控数据仿真中得到了验证。

关键词:故障诊断;类别不平衡;SMOTE;过采样技术
中图分类号:TP391.4 **文献标识码**:A **文章编号**:1673-629X(2016)09-0043-04
doi:10.3969/j.issn.1673-629X.2016.09.010

Fault Diagnosis Method Based on Weighted Synthetic Minority Oversampling Technique

HAN Zhi-yan,WANG Jian
(College of Engineering,Bohai University,Jinzhou 121000,China)

Abstract:The Synthetic Minority Oversampling Technique (SMOTE) is a famous oversampling method,whereas it doesn't consider the distribution of samples and latent noises in the data. In order to improve the performance of SMOTE,a modified method,the Weighted Synthetic Minority Oversampling Technique (WSMOTE),is proposed. WSMOTE introduces the neighborhood union to classify the samples into several groups,and different groups have different importance. Then,WSMOTE generates synthetic sample according to the different importance. The proposed method has a better performance when dealing with class imbalance data and it is demonstrated through its application to the semiconductor wafer fabrication process.

Key words:fault diagnosis;class imbalance;SMOTE;oversampling technique

0 引 言

近年来,半导体制造工业一直保持较高的增长速度。半导体制造是一个非常复杂的生产过程,由数百个步骤构成,其中晶元制造是其最关键的一步。晶元制造工艺包括一系列步骤,以在晶元表面覆盖特殊的材料层。在这个复杂的过程中,一些很小的缺陷就可以使最终的产品测试失败。因此,为了满足半导体工艺的质量要求,故障诊断与分类研究成为当前的热点问题[1]。

如今,随着数据收集和采集技术被广泛应用于半导体制造过程中,如何使用大量的已收集到的数据来有效地描述生产过程,极大地促进了基于数据驱动的故障诊断方法的研究工作。最近一些基于模式识别的故障诊断方法被提出以解决半导体制造过程中出现的非线性和多批次轨迹问题。例如,He 等[2]提出在半导体工业的故障检测中使用 k -最近邻(KNN)规则来完成故障分类。Verdier 等[3]同样应用了 KNN 规则,但他们提出的方法使用自适应马氏距离来代替传统的欧几里得距离。然而,在半导体故障诊断过程中的数据类别不平衡特性,给这些方法的应用带来了困难,由于与正常工况的数据相比,故障工况的数据常常难以获取,所以工业现场中收集的监测数据常常具有严重的类别不平衡特性。在这种情况下,传统的分类器倾向于将数据归类于多数类(正常工况),以得到更高的总体准确率而忽视了少数类(故障工况)的准确率。然而,在故障诊断中,最重视的往往是少数类(故障工

收稿日期:2015-10-28 **修回日期**:2016-02-24 **网络出版时间**:2016-08-23
基金项目:国家自然科学基金资助项目(61403042,61503038);辽宁省教育科研计划项目(L2013423)
作者简介:韩志艳(1982-),女,博士,副教授,研究方向为情感识别、语音识别。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160823.1359.042.html

况)的分类准确率。在解决这一问题的方法中,重采样技术最为常用,特别是合成少数类过采样技术(SMOTE)引起了研究者的广泛关注[4]。Chawla 的实验研究表明,SMOTE 能够比其他采样方法取得更好的效果[5]。该文在 SMOTE 的基础上,提出了一种加权合成少数类过采样技术(Weighted Synthetic Minority Oversampling Technic,WSMOTE),通过有选择的过采样少数类样本来平衡两类样本在数量上的差距。

1 合成少数类过采样技术

合成少数类过采样技术(the Synthetic Minority Oversampling Technique,SMOTE)是一种主要的过采样技术,主要用来解决在分类问题中出现的样本分布不均衡。该算法的思想是合成新的少数类样本,以获得均衡的样本分布。合成策略是对每个少数类样本 x,搜索 k 个少数类最近邻样本;若向上采样的倍率为 n,则在其 k 个最近邻样本中随机选择 n 个样本,记为 $y_1, y_2, \cdots, y_n$;在少数类样本 x 与 $y_j(j=1,2,\cdots,n)$ 之间随机线性插值,构造新的少数类样本 p_j。

$$p_j = x + \text{rand}(0,1) * (y_j - x) \quad (1)$$

其中,rand(0,1) 表示 (0,1) 内的一个随机数。

图 1 是一个 SMOTE 算法的范例。

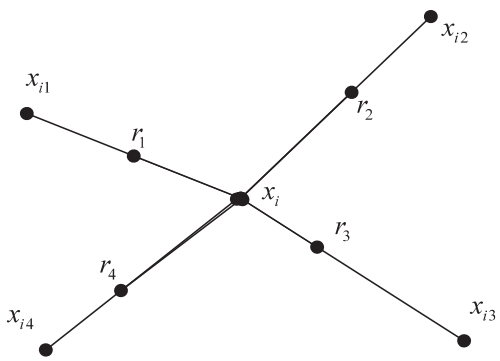

图 1 SMOTE 算法范例

如图所示:x_i 为某一个少数类样本,$x_{i1}, x_{i2}, x_{i3}, x_{i4}$ 分别为 x_i 的四个近邻,r_1, r_2, r_3, r_4 为生成的四个新的人造数据。

2 加权合成少数类过采样技术

SMOTE 是一种著名的过采样方法,但是它没有考虑样本的分布和潜在的噪声数据。为了改善 SMOTE 的性能,文中提出了加权合成少数类过采样技术(WSMOTE)。

由于基于流形假设的局部拓扑结构既受到类间的不平衡的影响又受到类内不平衡的干扰,因此 WSMOTE 算法分别从类内和类间两个层面研究样本的分布和潜在的噪声影响。在本节中,类间不平衡是指样本的多数类的数目不同于少数类的数目的情况;类内不平衡是指同一类样本是由许多不同的子群组成,而这些子群的重要性是不同的。

同 SMOTE 相似,WSMOTE 通过产生合成样本解决类间不平衡问题。在处理类内不平衡时,WSMOTE 通过引入邻域并将样本按照分布的不同划分为不同的组群再加权合成样本来解决。

如图 2 所示,点 q 和 r 分别是近邻的类间样本 x_q 和 x_r,$N(x_q)$ 和 $N(x_r)$ 是它们各自的近邻,其对应的邻域并写作 $N(x_q, x_r)$,其中 $N(x_q, x_r) = N(x_q) \cup N(x_r)$。显然,$x_q$ 和 x_r 的关系处于 $N(x_q, x_r)$ 的约束下。当 x_q 和 x_r 是类内近邻样本,邻域并也可以用同样的方式定义。

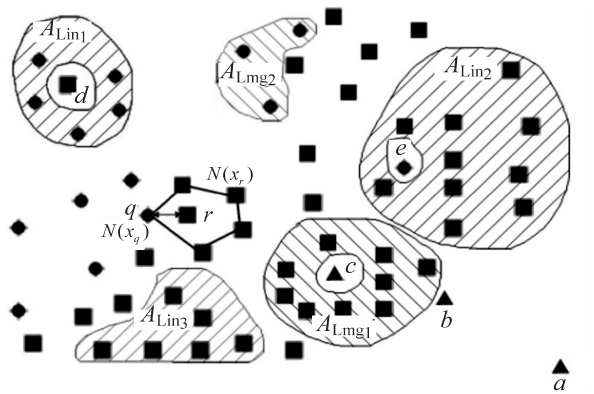

图 2 样本集划分的邻域并约束

在邻域并中,从局部类别分布上看,由于 $N(x_q, x_r)$ 对 x_q 和 x_r 间关系的约束能分解为 $N(x_q)$ 和 $N(x_r)$ 对 x_q 和 x_r 各自的约束。如果 $N_w(x_r) \neq \varnothing$ 且 $N_b(x_r) \neq \varnothing$,其中 $\varnothing$ 是空集,$N_w(x_r)$ 和 $N_b(x_r)$ 分别表示 x_r 的类内邻域和类间邻域,可以令 x_r 是一个边界样本。如果样本 x_r 的近邻都位于 $N_b(x_r)$ 里,即 $N_w(x_r) \neq \varnothing$,这样的样本可以假定是孤立样本。如果一个样本被同类近邻包围,即 $N_b(x_r) \neq \varnothing$,令 x_r 是内部样本。因此,根据局部类别分布与样本所属类别的数据量大小,样本可划分到六个不同子集中:

A_{Ny}:由大类和中等类的孤立样本所组成的噪声样本集;

A_{Lmg}:大类和中等类的边界样本集;

A_{Lin}:大类和中等类的内部样本集;

A_{Siso}:小类的孤立样本集;

A_{Smg}:小类的边界样本集;

A_{Sin}:小类的内部样本集。

在样本集中,每个样本仅仅属于一个集合,这六个子集的并集构成了整个样本集。图 2 给出了特征空间的一个场景示例,其中方块、圆块和三角形分别代表大类、中等类和小类的样本。样本 x_q 和 x_r 分别受 $N(x_q)$ 和 $N(x_r)$ 约束,x_q 和 x_r 之间的关系受 $N(x_q, x_r)$ 约束。根据样本子集的定义,样本可以如下归类:$A_{\text{Sin}} = \{a\}$,

$A_{Smg}=\{b\}$,$A_{Siso}=\{c\}$,$A_{Lin}=\{A_{Lin1}\cup A_{Lin2}\cup A_{Lin3}\cup\cdots\}$,$A_{Lmg}=\{A_{Lmg1}\cup A_{Lmg2}\cup A_{Lmg3}\cup\cdots\}$,并且$A_{Ny}=\{d,e\}$。

不同的局部分布类型对具有不同的采样价值。内部样本代表了一个特定类别的典型属性,所以可以看作标准样本。和内部样本不同,边界样本在特征空间中离类间样本很近,因此有更高的误分可能性。因为孤立样本与异类样本更相似,所以有最高的误分可能性。因此,WSMOTE 根据不同的策略选择生成合成样本。具体规则如下:算法随机地从 A_{Sin} 集合中选择样本的 k 近邻产生合成样本,从 A_{Smg} 集合中选择样本最近邻产生合成样本,对 A_{Siso} 集合不合成任何样本,移除 A_{Ny} 集合中的样本。

3　仿真实验及结果分析

文中使用 SECOM 数据集验证 WSMOTE 算法的有效性。首先介绍了不平衡数据分类性能的评估方法。然后,简要介绍了 SECOM 数据集。最后,分析了在 SECOM 数据集中获得的仿真结果。

3.1　不平衡数据分类性能评估方法

在故障诊断实践中,由于正常工况数据容易获得,而故障工况数据难以获得,导致训练数据广泛存在类不平衡情形[6-8]。当处理类分布不平衡数据时,由于多数类占优势,分类边界偏置于优势数据,经典分类算法面临对少数类预测能力下降的问题,从而影响整体预测性能。

表 1 所示的混淆矩阵表达了样例分类的分布情况。混淆矩阵是计算若干分类器性能度量的基础。

表 1　混淆矩阵

	预测正类	预测负类
实际正类	True Positive (TP)	False Negative (FN)
实际负类	False Positive (FP)	True Negative (TN)

对于两类问题,通常称少数类为正类,称多数类为负类,正确率 Acc 和错误率 Err 为:

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}} \tag{2}$$

$$\mathrm{Err}=\frac{\mathrm{FP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}} \tag{3}$$

正确率 Acc 和错误率 Err 是常用的分类器性能度量,但是,这两个度量对类不平衡敏感,过于偏置多数类。在处理不平衡数据时,使用 Acc 或 Err 将会导致性能比较的错误结果[9]。

以下度量由混淆矩阵派生,也是其他度量的基础:

真正率:

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{4}$$

真负率:

$$\mathrm{TNR}=\frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}} \tag{5}$$

假正率:

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \tag{6}$$

假负率:

$$\mathrm{FNR}=\frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} \tag{7}$$

显然,分类器想要在两个类别中均取得良好的分类性能,单靠其中某一个性能指标是不能胜任的,需要把其中某些指标结合起来,形成一种新的评价基准。

3.2　SECOM 数据集简介

文中使用的 SECOM 数据集[10]是从真实的半导体制造生产线上获取的相关数据。SECOM 数据集包含 2 个文件,数据文件包含 1 567 个样本,每个样本包含 591 个特征,标签文件包含每个样本的分类标签和采样时间。如同多数采自工业现场的数据,数据集中很多特征对应着空值或常值,这一情况需要在数据预处理阶段进行处理。

3.3　结果和分析

在数据预处理阶段,由于 SECOM 数据集中的某些特征包含空白值或常值,共删除了 137 个特征,这些特征符合 80% 的数据记录丢失或为常值,在剩余的 454 个特征中,使用 10 倍交叉验证技术验证用于比较的各种模型算法。所以,首先把 SECOM 数据集分成训练数据集和测试数据集,训练数据集包含从原始数据集中随机选择的 94 个故障样本和 1 037 个正常样本,测试数据集包含 250 个样本,其中,故障样本 104 个,正常样本 146 个。WSMOTE 中的 A_{Sin} 取值为 3。

为了比较 SMOTE+PCA(SPCA),WSMOTE+PCA(WPCA),SMOTE+FDA(SFDA),WSMOTE+FDA(WFDA),SMOTE+MFA(SMFA),WSMOTE+MFA(WMFA)的性能,在 SECOM 数据集分别使用它们进行特征选择,进行对比研究。其中,SPCA,SFDA 和 SMFA 是首先使用 SMOTE 进行类别数据再平衡后再和主元分析(Principal Component Analysis,PCA)[11-13]、费舍尔判别分析(Fisher Discriminant Analysis,FDA)[14]、边际费舍尔分析(Margin Fisher Analysis,MFA)[15]相结合产生的特征提取算法;WPCA,WFDA 和 WMFA 是首先使用 WSMOTE 进行类别数据再平衡后再和 PCA,FDA 和 MFA 相结合产生的特征提取算法。图 3 分别比较了六种算法的多种性能指标。

从图 3 可以看出,在六种算法中,WFDA 拥有最佳的分类性能,因为它能够满足对一个好的特征选择算法的期望,即拥有高的 TPR,TNR 和 Acc,拥有低的 FPR 和 FNR。而且,所有使用了 WSMOTE 算法的特

征选择方法在故障样本的识别性能上均优于使用 SMOTE 算法的特征选择方法。它表明,WSMOTE 算法可以通过有选择地增加故障样本的数量,改进训练数据集的样本多样性,从而改善特征选择算法的性能。但是,有时使用 WSMOTE 算法的模型会降低多数类(正常样本)的分类性能,这是由于想在两个类别中同时获得更优的性能是一件困难的事情,因此在实施这一算法时应综合考虑多方面因素。

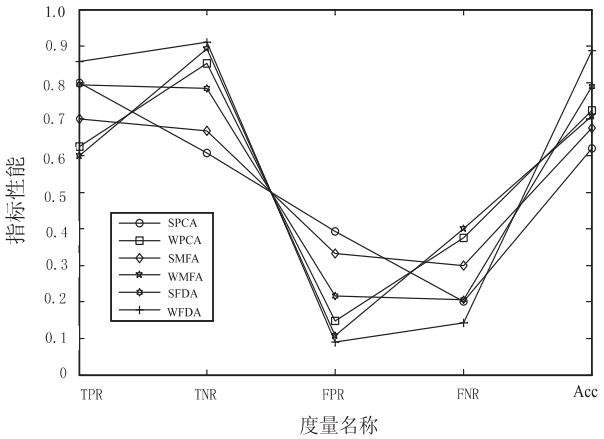

图 3 六种算法的性能指标图

4 结束语

在 SMOTE 的基础上,提出 WSMOTE 算法用于解决故障诊断过程中因故障数据难以获得而出现的数据类别不平衡问题。该算法分别从类内和类间两个层面研究样本的分布和潜在的噪声影响。同 SMOTE 相似,WSMOTE 通过产生合成样本解决类间不平衡问题。在处理类内不平衡时,WSMOTE 通过引入邻域并将样本按照分布的不同划分为不同的组群,不同的群组拥有不同的采样价值,然后根据采样价值的不同加权合成样本来解决。WSMOTE 在处理类别不平衡数据时具有优异的性能,并在半导体制造过程的监控数据仿真中得到了验证。

参考文献:

[1] Bleakie A, Djurdjanovic D. Feature extraction, condition monitoring, and fault modeling in semiconductor manufacturing systems[J]. Computers in Industry, 2013, 64(3): 203-213.

[2] He Q P, Wang J. Fault detection using the k-Nearest neighbor rule for semiconductor manufacturing processes[J]. IEEE Transactions on Semiconductor Manufacturing, 2007, 20(4): 345-354.

[3] Verdier G, Ferreira A. Adaptive mahalanobis distance and k-nearest neighbor rule for fault detection in semiconductor manufacturing[J]. IEEE Transactions on Semiconductor Manufacturing, 2011, 24(1): 59-68.

[4] Chawla N V, Hall L O, Bowyer K W, et al. SMOTE: synthetic minority over sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[5] Chawla N V. C4.5 and imbalanced datasets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure[C]//Proceedings of the workshop on learning from imbalanced datasets. Washington D C: [s. n.], 2003: 17-23.

[6] Chawla N V. Data mining and knowledge discovery handbook[M]. Berlin: Springer, 2010: 857-886.

[7] 王和勇,樊泓坤,姚正安. SMOTE 和 Biased-SVM 相结合的不平衡数据分类方法[J]. 计算机科学, 2008, 35(5): 174-176.

[8] Cebe M, Gunduz-Demir C. Qualitative test-cost sensitive classification[J]. Pattern Recognition Letters, 2010, 31(13): 2043-2051.

[9] Elazrneh W, Japkowicz N, Matwin S. Evaluating misclassifications in imbalanced data[C]//Proc of the 17th European conference on machine learnin. Berlin: Springer, 2006: 126-137.

[10] McCann M, Li Y, Maguire L. Causality challenge: benchmarking relevant signal components for effective monitoring and process control[C]//Proc of JMLR. Canada: [s. n.], 2008: 277-288.

[11] Wang T, Xu H, Han J, et al. Cascaded h-bridge multilevel inverter system fault diagnosis using a PCA and multiclass relevance vector machine approach[J]. IEEE Transactions on Power Electronics, 2015, 30(12): 7006-7018.

[12] Ding S, Zhang P, Ding E, et al. On the application of PCA technique to fault diagnosis[J]. Tsinghua Science and Technology, 2010, 15(2): 138-144.

[13] Wang N, Yuan Z H, Wang D. Improving process fault detection and diagnosis using robust PCA and robust FDA[C]//Proc of WRI world congress on computer science and information engineering. USA: IEEE, 2009: 54-59.

[14] Tang X C, Yuan L. Monitoring and fault diagnosis using fisher discrimnant analysis[C]//Proc of the international conference on machine learning and cybernetics. USA: IEEE, 2007: 1100-1105.

[15] Tsang I W, Kocsor A, Kwok J T Y. Large-scale maximum margin discriminant analysis using core vector machines[J]. IEEE Transactions on Neural Networks, 2008, 19(4): 610-624.