一种改进的衡量数据不平衡的标准

## Abstract

不平衡分类是违反样本均匀分布假设的分类问题，在这类问题中，感兴趣的样本识别率较低。传统的不平衡数据集都按照样本数的不平衡率来衡量的，但是在各种研究中，分类结果不只与样本数量相关，样本分布在其中起到了更重要的作用，本文中提出了一种改进的泛化不平衡率，较好地表示了样本的分布情况与分类结果的关系，实验结果显示，该衡量标准和分类结果具有更高的相关性，更好地衡量数据集的分类难度。

分类问题是机器学习中非常重要的一部分，在传统的分类问题中，模型训练是基于样本分布均匀的假设，因而每个样本的分类代价是一致的。但是在现实的数据集中，样本均匀分布假设难以符合，在这类问题中，由于感兴趣的类别中样本数量过少，为了追求全局准确率，传统分类器容易忽视该类样本的识别，造成该类别识别率很低。不平衡分类问题出现在许多领域中，例如生物信息学[1, 2]、遥感图像识别[3]、网络安全中的隐私保护[4]等。不平衡问题的覆盖范围广，具有十分重要的现实意义。

正是因为样本数量对分类结果造成了显而易见的影响，所以样本数量不平衡率作为不平衡数据集的衡量指标，已经流行了多年

由于长期使用样本size的不平衡度来作为数据集的不平衡度，并基于此不平衡度来指导采样过程中的终止与否标准，因而不平衡度对于采样算法来说，有着非常重要的作用，但是样本size的不平衡度并不足以衡量一个数据集的分类难易程度，研究表明，样本数量在足够多的情况下，即相对稀少并不会引起少数类分类效果的降低，而是在样本数量严重不足的情况下，样本数量不平衡才会对分类结果产生影响。

因此，本课题从样本分类难度的不一致考虑数据集的不平衡程度，改进传统的基于样本集合大小的不平衡率计算方法，提高不平衡率和最终分类结果的相关度。

## Introduction

分类问题是机器学习中非常重要的一部分，在传统的分类问题中，模型训练是基于样本分布均匀的假设，因而每个样本的分类代价是一致的。但是在现实的数据集中，一方面我们对不同类别的感兴趣程度不一样，另一方面不同类别的数据分类难度也不相同，因而出现了不平衡分类问题：样本分布违背均匀分布假设，造成人们感兴趣的类别识别效果较差。在这类问题中，传统的机器学习方法是基于每类样本的分类代价相同的假设，由于少数类数量过少，为了追求全局准确率，传统分类器容易忽视少数类样本的识别，造成少数类识别率很低。例如在医疗健康体检数据中，绝大多数的体检报告结果是正常的，只有少部分的体检报告是真实需要治疗的，在这些少数类中，如果产生误诊断为正常，则会对病人的权益产生极大的损害，甚至错过最佳救治时间，但是如果为了保证较低的误诊率而将大多数体检报告诊断为有病，则会浪费很多医疗资源，甚至影响医患关系。在这一类的情况下，不同类别的分类代价不同，而我们感兴趣的类别识别难度较高，即构成不平衡分类问题。不平衡分类问题出现在许多领域中，例如生物信息学[1, 2]、遥感图像识别[3]、网络安全中的隐私保护[4]等。不平衡问题的覆盖范围广，具有十分重要的现实意义。

不平衡分类具有两个特征：样本数量分布不均衡；样本误分类代价不一致。为了解决这两个问题，学者们提出了很多方法：数据层面和算法层面。在数据层面的算法中，利用数据重采样技术使数据达到平衡状态[5–8]。这类算法中，指代平衡状态的指标是非常重要的，不同指标下的平衡状态通常有着不同的样本分布。传统的数据集不平衡度都是考虑样本数量不平衡，而忽略了样本分布对于其分类结果的影响。算法层面的算法则是针对不平衡数据集的第二个特征，主要有两大类算法：代价敏感[9]：对不同类别的样本的分类代价予以加权，以提高受关注的类别的分类效果；集成算法[10]：通过不同的分类器的集成，尽可能区分出不同类别的样本，提高最终的分类效果。

由于长期使用样本size的不平衡度来作为数据集的不平衡度，并基于此不平衡度来指导采样过程中的终止与否标准，因而不平衡度对于数据层面的算法来说，有着非常重要的作用，但是样本size的不平衡度并不足以衡量一个数据集的分类难易程度，研究表明，样本数量在足够多的情况下，即相对稀少并不会引起少数类分类效果的降低，而是在样本数量严重不足的情况下，样本数量不平衡才会对分类结果产生影响；同时，学者们还分析了其他很多影响分类结果的因素，例如类内不平衡：指的是在同一类别中，出现了同类别的样本聚集成不同的簇的情况，分类器容易将大簇所在的区域划分为该样本，而将小簇识别为噪声样本或者是别的类别的样本；边界重叠问题，由于全局准确度的定义问题，基于边界分类器在划分边界时，容易将少数类样本划分为多数类样本，从而造成少数类识别率下降。

因此，本课题从样本分类难度的不一致考虑数据集的不平衡程度，改进传统的基于样本集合大小的不平衡率计算方法，提高不平衡率和最终分类结果的相关度。

## 相关工作

不平衡衡量指标可以分为两种：局部指标和全局指标，局部指标指的是遍历数据集中的每个样本，计算该样本的某个指标，通常伴随着k-nn算法，整个样本集合的分类难度则以该数据集中的全部样本的指标均值定义，因为这类指标中，包含了对每个样本的计算，因而可以用于采样算法中，以衡量新数据集的分类难度；全局指标：指的是针对整个数据集中的样本计算得来的某个结果，或者是统计分析中得出的各种指标，通常伴随着正负类子集的指标结果的各类计算，这类计算难以在单一样本上实现，因而只能作为数据集的衡量指标，难以在采样算法中发挥作用，因为单一样本的移动很难对原始的指标结果产生影响。

不平衡衡量指标的作用有两个：衡量样本分类难度；在重采样方法中作为indicator数值使用，因此，为了较好的完成该作用，该指标应当具有比较高的和分类结果的相关性。

目前有很多计算样本分类难度的指标，以下为比较经典的指标：

以二分类数据集为例，给定数据集为X，其中包含P类样本（少数类样本）N+个，N类样本（多数类样本）N-个，总样本数为N

f1[11]

A classical measure of the discriminative power of the covariates,or features, is Fisher’s discriminant ratio:

|  |  |  |
| --- | --- | --- |
|  |  | (2-1) |

where are the means and variances of the two classes, respectively. For multidimensional problems, it is not necessarily the case that all features contribute to class discrimination, so the maximum f over all features can be used. However, a zero maximum f does not necessarily mean that the classes are not separable, as it could just be that the separating boundary is not parallel to an axis in any of the given features.

ir

Imbalance ratio the definition is as follows, it is defined as the size sample ratio:

|  |  |  |
| --- | --- | --- |
|  |  | (2-1) |

As the size has less importance than the sample distribution, so the improvements in the imbalance evaluation are mostly related to the sample distribution, when the data is 同分布的情况下，sample size是能够体现样本集合分类难度的，但是在图（） 中，图a中数据的不平衡度为4，而b中为1，但是(a)中两个类别有清晰的线性边界，但是在同样的线性模型下，(b)中的样本却无法得到100%的分类结果。因此实际上(b)中的数据集比(a)中的更难分类，这与IR的比较结果相反，由于IR是样本size的比例，而不包含任何样本分布信息，因此IR中无法体现数据集本身分布的复杂程度

|  |  |
| --- | --- |
| F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\figa.png | F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\figb.png |
| (a) | (b) |

cm[12]

Our approach focuses on the local information for each data point via the nearest neighbors, and uses this information to capture data complexity. Calculation involves finding the k nearest neighbors of every data point in the class, where k is odd. If the majority of the neighbors are in the same class, this point is designated as easy to classify; if most of the neighbors are in the opposing class, it is difficult. Complexity can be measured as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (2-2) |

Where I(.) is the indicator function. The overall measure is

|  |  |  |
| --- | --- | --- |
|  |  | (2-2) |

Cm每个样本的分类难度由其近邻的种类决定，即如果某个样本的近邻中包含越多的同类样本，则该样本的分类难度越低，相反，如果某个样本都是被其他类别的样本包围，则分类器很难分类正确，利用样本的近邻中包含的不同类别的样本个数的均值作为该样本的分类难度，cm越高，则样本集合分类难度越大，cm可以有效的反映出数据集的分类难度。

Gir[13]

The gir is an improvement of cm, it focus on the difference in the difficulty of classifying the samples in different classes, a dataset with a larger gir is more difficult to get a good performance of the minority class, as the classifier tends to be trained with the easier samples according to the Occam shaver principle, because we tend to use a as simple as possible classifier to fit the whole dataset, while the more difficult samples need a more complex classifier, which may cause overfitting with single classifier, so this is the reason why ensemble can be effective in the imbalance classification, as they have the different classifiers corresponding to different levels of sample classification difficulty.

|  |  |  |
| --- | --- | --- |
|  |  | (2-2) |

其中Ir（x，X）是一个指示函数，对于某个样本x，如果其k近邻中，该近邻的类标和x相同则为1，不同则为0

|  |  |  |
| --- | --- | --- |
|  |  | (2-4) |

|  |  |  |
| --- | --- | --- |
| (2-1) |  | (2-6) |

Gir考虑到了正负类子集中的分类难度不一致，对cm进行改进，gir为正负类样本子集的cm之差，并成功将gir应用到过采样和欠采样的算法中，实验结果表明，基于gir的重采样算法可以有效提高the classification performance.

## 本文方法

在GIR的定义过程中，每个样本的GIR是计算其k近邻中的同类别的平均样本个数，但是在分类过程中，k近邻的距离同样会对分类结果造成影响，因为前n个同类标和后n个同类标对样本的类别的影响显然不是一样的；其次，数据集的GIR是通过负类-正类来进行计算的，但是gir是一个相对指标，例如表中所示，从最终的结果来看，两个数据集的GIR结果相同，但是显然后者的分类难度比前者要高，因此gir并不足以完全indicate数据集的分类复杂度。

|  |  |  |
| --- | --- | --- |
|  | (a) | (b) |
|  | 0.9 | 0.5 |
|  | 0.7 | 0.3 |
|  | 0.2 | 0.2 |

因为传统的不平衡率是计算少数类和多数类的样本size的比率，而忽略了样本的分布问题，而样本分布在分类结果中起到了比较重要的作用，因此考虑了样本分布的指标，则会有比较好的指示分类结果的作用。基于这一思想，本文中提出了一种改进的GIR，称之为IGIR，该指标的动机为如果样本周围的同类标样本数较多，则该样本容易被分类，而样本周围的同类标样本数较少，则样本难以被分类。不同距离的同类标样本对该样本的分类结果的作用不同。

在计算IGIR过程中，首先计算数据集中每个样本的k近邻，并保留其近邻的类标，根据公式2中的计算方式，k近邻的权重逐级较低直到0。不采用距离的原因主要是不同的样本的近邻的距离不一致，会导致在计算过程中，每个样本的近邻权重都不一致，因而无法对整体的结果有个比较标准；其次，为了合理描述数据集的分类难度，我们修正了整个数据集的GIR的计算方法，对不同类别的样本集合中的样本分类难度进行统计，igir定义为不同类别的样本的分类难度相乘，在这种情况下，为了保证其数量级不变，对其进行开方处理，从而避免了原来的GIR中的相对性，以便更好地对数据集的分类难度进行衡量。

weight-=1/k

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | (2-3) |
|  | | |  | (2-5) | | |
|  | |  | | | (2-9) | |

Igir可以理解为是样本在加权的knn下，正类和负类样本的分类准确概率的平均值，即样本中的相同类标的近邻数越多，该样本越可能被分类为原始类标，即越大可能分类正确，分类难度越低；相反则越小可能分类正确，分类难度越高。

|  |
| --- |
| Algorithm 1：IGIR：A new measurement of the imbalanced dataset. |
| Input: : dataset with N samples, consisting of positive samples in P and negative samples in N, Y: the corresponding labels.  Output: the measurement IGIR of X U Y.  Procedure:   1. For each x in X with label 2. Compute the k nearest neighbors of x 3. Count the number of samples with the same label and save the corresponding weight 4. calculate the I(x) for sample x 5. Calculate the average , according to the yighted-T\_, weighted\_d\_tares, and the SSE represents the error sum of squares.ult samples need a more complex classif formula() 6. Calculate the overall IGIR according to the formula() |

## 实验结果及验证

Dataset

文中实验数据来自于UCI机器学习数据库[14]，有些是多类标数据集，为了获得高的不平衡率，我们选中其中某一个类别为少数类，其余的样本均归为多数类。本文中的分类结果为C4.5分类器的十次十折交叉验证的平均值。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Datasets | Samples | Attributes | Target | Minority |
| breasttissue | 106 | 9 | Carcinoma | 21 |
| breastw | 699 | 9 | malignant | 241 |
| diabetes | 768 | 8 | 0 | 268 |
| german | 1000 | 24 | 2 | 300 |
| glass | 214 | 9 | 1 2 3 | 51 |
| haberman | 306 | 3 | 1 | 81 |
| ionosphere | 351 | 34 | G | 126 |
| movement | 360 | 90 | 1 | 24 |
| satimage | 6435 | 36 | 4 | 703 |
| segment-challenge | 1500 | 19 | brick face | 205 |
| sonar | 208 | 60 | R | 97 |
| spect | 267 | 22 | 1 | 55 |
| vehicle | 846 | 18 | van | 199 |
| vertebral | 310 | 6 | AB | 100 |
| wpbc | 198 | 33 | 1 | 47 |
| yeast0 | 1484 | 8 | 0 | 244 |
| yeast1 | 1484 | 8 | 1 | 429 |
| yeast2 | 1484 | 8 | 2 | 463 |
| yeast6 | 1484 | 8 | 6 | 163 |

Evaluation

在不平衡数据二分类中，常常使用混淆矩阵来评估分类器的性能，其定义如下：

表2-4混淆矩阵

|  |  |  |
| --- | --- | --- |
|  | Positive prediction | Negative prediction |
| Positive class | True positive(TP) | False negative(FN) |
| Negative class | False positive(FP) | True negative(TN) |

其中，代表正类被正确预测为正类的样本个数，TN则是负类被正确预测为负类的样本数，FN则表示正类被错误判定为负类的样本数，而FP则为负类被错误判定为正类的样本数，目前出现了一些新的不平衡数据的分类评价指标，例如F-value 和 G-mean等方法[15]。因此在实验结果分析中，采用了F-value和等能够分析整体分类情况的预测指标，同时，出于对少数类的更加关注，还对比了少数类的分类准确度sensitivity。

|  |  |  |
| --- | --- | --- |
|  |  | (2-22) |
|  |  | (2-22) |
|  |  | (2-22) |
|  |  | (2-23) |

F-value 是衡量准确率和召回率的分类评价指标，本文中计算了F1-min，比较偏向对少数类的分类性能评价，表示少数类分类精度和多数类分类精度的几何平均值，只有在少数类和多数类分类精度同时都高的情况下，此时，的值最大。

实验设置：

F1\_min中β=1，所有涉及knn的指标计算中k=5，分类器为C4.5，所有结果均为10次十折交叉验证的平均值。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ir | gir | cm | f1 | igir | F1\_min | gmean | Sensitivity |
| breasttissue | 4.05 | 0.30 | 0.24 | 3.33 | 0.46 | 0.78 | 0.83 | 0.83 |
| breastw | 1.90 | 0.05 | 0.05 | 3.47 | 0.57 | 0.91 | 0.92 | 0.88 |
| diabetes | 1.87 | 0.22 | 0.49 | 0.58 | 0.37 | 0.57 | 0.66 | 0.57 |
| german | 2.33 | 0.37 | 0.52 | 0.35 | 0.33 | 0.47 | 0.60 | 0.48 |
| glass | 3.20 | 0.19 | 0.12 | 3.31 | 0.53 | 0.75 | 0.80 | 0.75 |
| haberman | 2.78 | 0.43 | 0.46 | 0.18 | 0.32 | 0.24 | 0.35 | 0.30 |
| ionosphere | 1.79 | 0.39 | 0.21 | 0.61 | 0.46 | 0.84 | 0.87 | 0.82 |
| movement | 14.00 | 0.44 | 0.06 | 0.98 | 0.47 | 0.59 | 0.77 | 0.69 |
| satimage | 8.15 | 0.04 | 0.01 | 5.01 | 0.59 | 0.95 | 0.97 | 0.94 |
| Segment\* | 6.32 | 0.04 | 0.03 | 1.81 | 0.58 | 0.97 | 0.98 | 0.97 |
| sonar | 1.14 | 0.16 | 0.34 | 0.46 | 0.47 | 0.59 | 0.58 | 0.62 |
| spect | 3.85 | 0.61 | 0.33 | 0.60 | 0.32 | 0.50 | 0.65 | 0.51 |
| vehicle | 3.25 | 0.10 | 0.12 | 1.12 | 0.54 | 0.88 | 0.92 | 0.89 |
| vertebral | 2.10 | 0.14 | 0.31 | 0.75 | 0.47 | 0.67 | 0.71 | 0.66 |
| wpbc | 3.21 | 0.39 | 0.43 | 0.47 | 0.32 | 0.42 | 0.56 | 0.46 |
| yeast0 | 5.08 | 0.38 | 0.21 | 0.74 | 0.41 | 0.49 | 0.68 | 0.53 |
| yeast1 | 2.46 | 0.34 | 0.42 | 0.24 | 0.37 | 0.48 | 0.62 | 0.50 |
| yeast2 | 2.21 | 0.26 | 0.48 | 0.21 | 0.37 | 0.49 | 0.61 | 0.49 |
| yeast6 | 8.10 | 0.33 | 0.08 | 2.75 | 0.47 | 0.69 | 0.82 | 0.71 |

注：segment 是 segment-challenge简称

|  |  |
| --- | --- |
| F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\imbalance-ratio.png | F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\gir.png |
| (a) | (b) |
| F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\f1.png | F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\cm.png |
| (c) | (d) |
| F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\wei-igir.png | |
| (e) | |

以少数类的分类准确度为例，图显示了不同衡量指标和分类结果间的关系，从图中可以看出，cm和igir作为统一数据集的评价指标，跟少数类分类准确度的相关性呈现线性趋势，其余的没有明显的趋势。此外，cm中的点更分散而igir中的更集中。

原因分析：因为在计算IGIR中，我们计算每个样本的k近邻的平均同类标样本的个数，因此计算的值可以看做是该样本分类为自己相同类别的概率，在一定程度上可以反映该样本的分类难易程度，将不同类别的样本的分类难易程度的均值相乘，在一定程度上可以视为这个数据集在knn分类器下的gmean，对于其他分类器的分类效果也具有一定的指示效果。

为了定量分析不同指标同分类结果的关系，本文中使用决定系数对结果进行进一步分析。反映了y的波动有多少百分比能被x的波动所描述，即表征依变数Y的变异中有多少百分比,可由控制的自变数X来解释.

Where , SST represents the total sum of squares, SSR represents the total sum of square, SSR represents the regression sum of squares, and the SSE represents the error sum of squares.

表中的的统计学指标也显示了igir的优越性，本文中提出的igir更能indicate分类结果，对数据集的分类难度和重采样的指示具有更好的效果。

|  |  |  |  |
| --- | --- | --- | --- |
|  | Fmeasure | gmean | sensitivity |
| igir | 0.92 | 0.88 | 0.93 |
| ir | 0.18 | 0.34 | 0.28 |
| gir | -0.70 | -0.58 | -0.67 |
| cm | -0.80 | -0.85 | -0.84 |
| f1 | 0.70 | 0.70 | 0.71 |

## 结论

本文中提出了一种改进的不平衡数据集衡量方法，该指标与分类结果具有更高的相关性，能更好的指导重采样算法。未来的工作将会基于此指标，进行重采样算法的研究，并希望能够提高分类结果。

## 参考文献

1. Wang Y, Li X, Tao B (2016) Improving classification of mature microRNA by solving class imbalance problem. Sci Rep 6:25941

2. Stegmayer G, Yones C, Kamenetzky L, Milone DH (2016) High class-imbalance in pre-miRNA prediction: a novel approach based on deepSOM. IEEE/ACM Trans Comput Biol Bioinform 1–1

3. Leichtle T, Geiß C, Lakes T, Taubenböck H (2017) Class imbalance in unsupervised change detection – A diagnostic analysis from urban remote sensing. Int J Appl Earth Obs Geoinformation 60:83–98

4. Li C, Liu S (2018) A comparative study of the class imbalance problem in Twitter spam detection. Concurr. Comput. Pract. Exp. 30:

5. Wang Q (2014) A Hybrid Sampling SVM Approach to Imbalanced Data Classification. Abstr Appl Anal-6-11 2014:1–7

6. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

7. Liu XY, Wu J, Zhou ZH (2009) Exploratory Undersampling for Class-Imbalance Learning. IEEE Trans Syst Man Cybern Part B 39:539–550

8. Savitsky TD, Srivastava S (2018) Scalable Bayes under Informative Sampling. Scand. J. Stat.

9. Zhang C, Wang G, Zhou Y, Jiang J (2017) A New Approach for Imbalanced Data Classification Based on Minimize Loss Learning. In: IEEE Second Int. Conf. Data Sci. Cyberspace. pp 82–87

10. Provost F (2008) Machine Learning from Imbalanced Data Sets 101 (Extended Abstract). In: Soft Comput. Pattern Recognit. SoCPaR 2011 Int. Conf. Of. pp 435–439

11. Ho TK (2002) A Data Complexity Analysis of Comparative Advantages of Decision Forest Constructors. Pattern Anal Appl 5:102–112

12. Anwar N, Jones G, Ganesh S (2014) Measurement of data complexity for classification problems with unbalanced data. Stat Anal Data Min Asa Data Sci J 7:194–211

13. Tang B, He H (2017) GIR-based Ensemble Sampling Approaches for Imbalanced Learning. Pattern Recognit. 71:

14. Amini MR, Usunier N, Goutte C (2013) uci-dataset-url.

15. Menardi G, Torelli N (2014) Training and assessing classification rules with imbalanced data. Data Min Knowl Discov 28:92–122

16. Zou Q, Xie S, Lin Z, Wu M, Ju Y (2016) Finding the Best Classification Threshold in Imbalanced Classification. Big Data Res 5:2–8