Dealing With Imbalanced Datasets

104403553 Lin, Ting-Wei, 104403555 Yang, Jhe-Sheng

Abstract

Imbalanced dataset problems are some of the most frequently discussed topics when it comes to data mining. The importance of the information provided by minority data cannot be overlooked, since oftenly some of the most critical pieces of details can only be revealed by them. The consequences of using imbalanced when doing data training without proper preparations of the datasets may be severe. This report focuses on introducing imbalanced datasets, tactics when encountering imbalanced datasets, and a brief analysis of algorithms performance when dealing with imbalanced data. Examples of imbalanced datasets as well as experiments conducted on the data are also featured. The results will then be analysed and compared with not-so-imbalanced datasets.

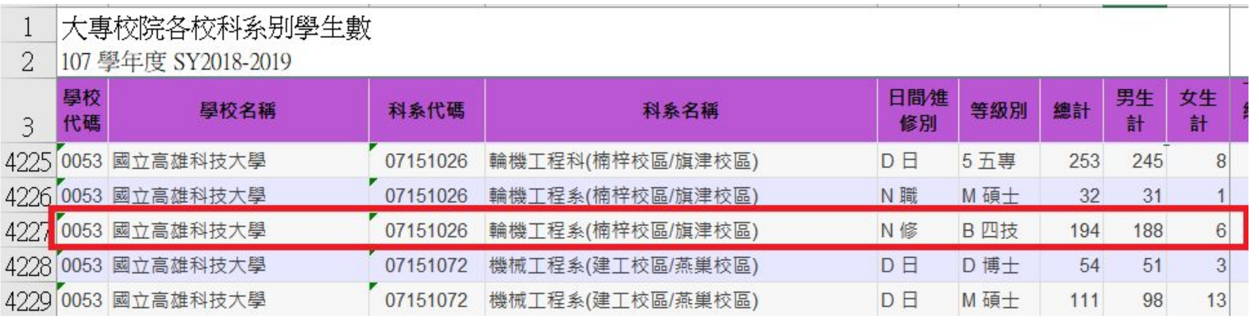
1.What is an Imbalanced dataset

An imbalanced dataset refers to datasets with data points distributed severely unevenly between classes. Slight inequality of number of data points(60:40) in different classes may not cause severe damage to the model trained.[1] An imbalanced dataset doesn't imply it's abnormal because data collected are sometime skewed, such suicide rate, crime rate and so on. If a dataset is highly imbalanced, for example a dataset which have are two classes with a ratio of 90:10, can cause a critical problem. The problem caused by an imbalanced dataset will be introduced in Section 2.

2.A brief example

We used "大專校院各校科系別學生數" provided by 教育部統計處 to introduce the concepts of imbalanced dataset.

Generally, male students are more encouraged to engaged in Science or Engineering studies, and that is no exception in Taiwan. After roughly examining the data, we can find some departments have an imbalanced gender proportion, and we chose the most significant one to introduce.



1 大專校院各校科系別學生數

2 107 學年度 SY2018-2019

3	學校代碼	學校名稱	科系代碼	科系名稱	日間/進修別	等級別	總計	男生計	女生計
4225	0053	國立高雄科技大學	07151026	輪機工程科(楠梓校區/旗津校區)	D 日	5 五專	253	245	8
4226	0053	國立高雄科技大學	07151026	輪機工程系(楠梓校區/旗津校區)	N 職	M 碩士	32	31	1
4227	0053	國立高雄科技大學	07151026	輪機工程系(楠梓校區/旗津校區)	N 修	B 四技	194	188	6
4228	0053	國立高雄科技大學	07151072	機械工程系(建工校區/燕巢校區)	D 日	D 博士	54	51	3
4229	0053	國立高雄科技大學	07151072	機械工程系(建工校區/燕巢校區)	D 日	M 碩士	111	98	13

Fig 1. Gender statistics of 高雄科技大學輪機工程學系大學部 [2]

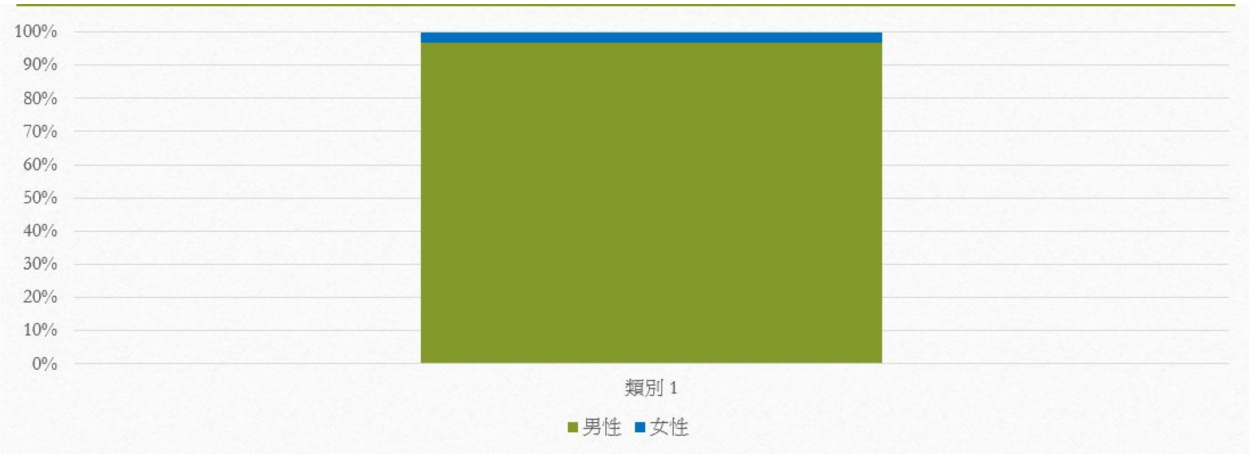
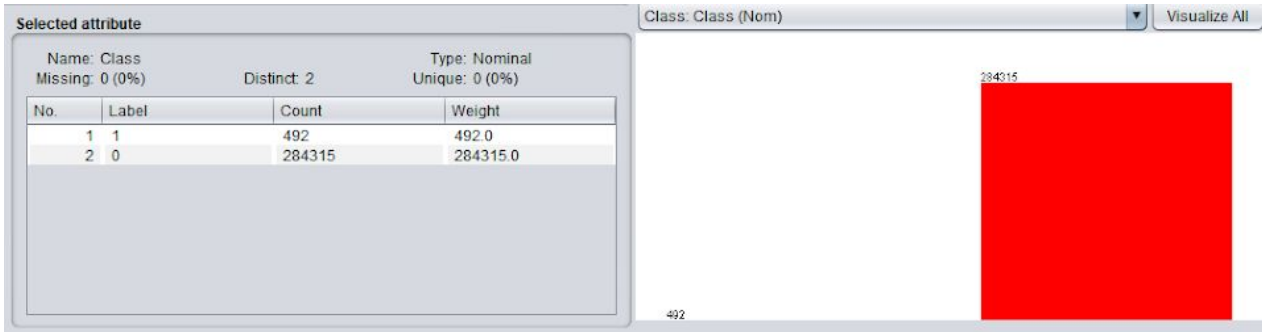


Fig 2. Gender statistics of 高雄科技大學輪機工程學系大學部

We can see that there is a huge difference between the proportion of male and female(188:6)in this department through this figure. If someone decided to use this dataset as a sample for research, we can say that he/she has used an imbalanced data set.We introduced the basic concept of an imbalanced data set in previous section, and we will show a brief example of a severely imbalanced dataset and the consequences of not handling it carefully.

A typical imbalanced dataset can be found in Kaggle named Credit Card Fraud Dataset.[3] This dataset is a dataset which has a binary (2-class) classification, class 1 means it's a fraudulent transactions, 0 otherwise. However in our daily life, most of transactions are normal transactions, and only few of them are fraudulent. In this dataset, the amount between normal transactions and fraudulent transactions are 284315 and 492, and the rate between them are 99.828%:0.172%.

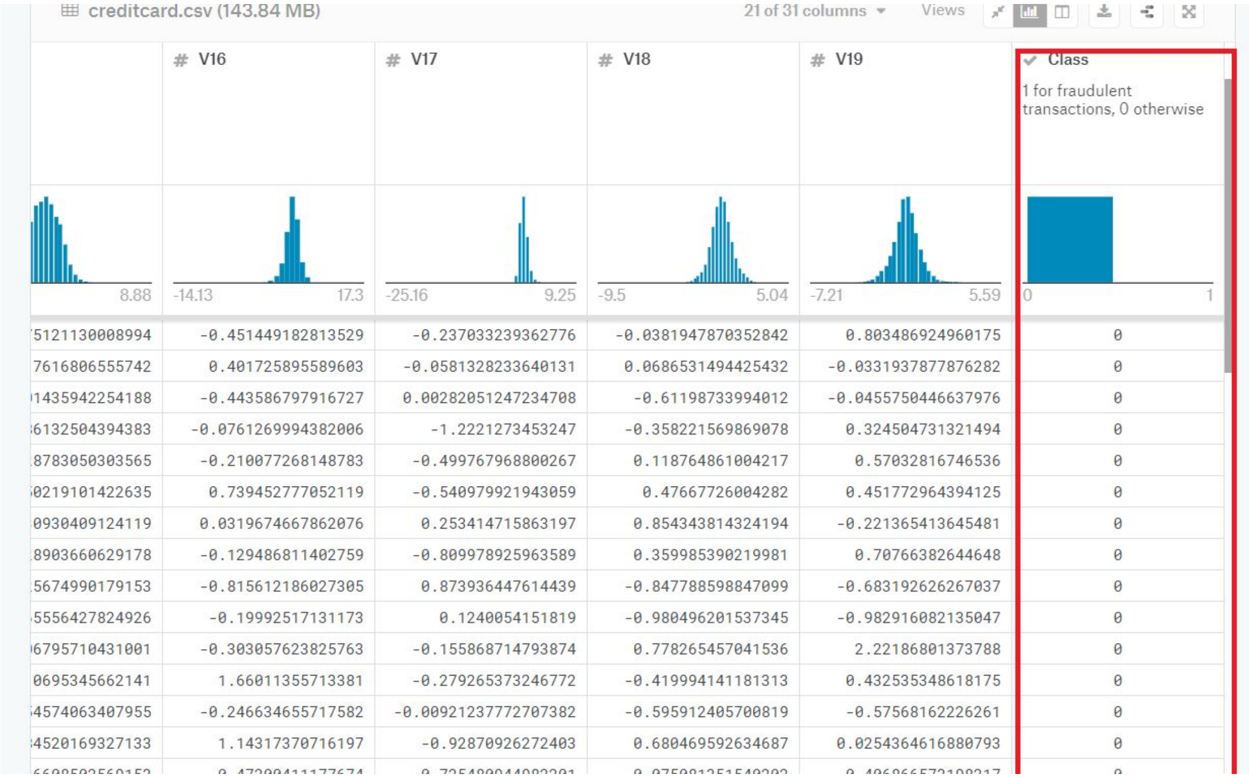


Selected attribute

Name: Class
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	1	492	492.0
2	0	284315	284315.0

Fig 3. The class bar graph of Credit Card Fraud Dataset



	# V16	# V17	# V18	# V19	Class 1 for fraudulent transactions, 0 otherwise
5121130008994	-0.451449182813529	-0.237033239362776	-0.0381947870352842	0.803486924960175	0
7616806555742	0.401725895589603	-0.0581328233640131	0.0686531494425432	-0.0331937877876282	0
1435942254188	-0.443586797916727	0.00282051247234708	-0.61198733994012	-0.0455750446637976	0
6132504394383	-0.0761269994382006	-1.2221273453247	-0.358221569869078	0.324504731321494	0
8783050303565	-0.210077268148783	-0.499767968800267	0.118764861004217	0.57032816746536	0
0219101422635	0.739452777052119	-0.540979921943059	0.47667726004282	0.451772964394125	0
0930409124119	0.0319674667862076	0.253414715863197	0.854343814324194	-0.221365413645481	0
8903660629178	-0.129486811402759	-0.809978925963589	0.359985390219981	0.70766382644648	0
5674990179153	-0.815612186027305	0.873936447614439	-0.847788598847099	-0.683192626267037	0
5556427824926	-0.19992517131173	0.1240054151819	-0.980496201537345	-0.982916082135047	0
6795710431001	-0.303057623825763	-0.155868714793874	0.778265457041536	2.22186801373788	0
0695345662141	1.66011355713381	-0.279265373246772	-0.419994141181313	0.432535348618175	0
4574063407955	-0.246634655717582	-0.00921237772707382	-0.595912405700819	-0.57568162226261	0
4520169327133	1.14317370716197	-0.92870926272403	0.680469592634687	0.0254364616880793	0

Fig 4. Overview of Credit Card Fraud Dataset

The last column of this dataset is the class column. Unfortunately, we can't get the detail information of other columns due to confidentiality issues.

The model achieved from training using this dataset might be faultful. We can get a 98% accuracy if all the guesses are normal transactions, therefore creating a paradox where the accuracy is unbelievably high and yet the model doesn't actually mean anything. Therefore, means have to be developed in order to get undesired results. In next section, we will show methods to deal with this problem.

3. How to deal with imbalanced datasets

Methods for dealing with problems concerning imbalanced datasets mainly focus on the data level, with several other targeting other aspects, eg. training algorithms, or the metric for evaluating the performance of the model. In this section, we will be discussing the methods for handling imbalanced datasets.

3.1 Collect more data

Collect more data is one of the most overlooked method. A larger dataset might reveal a more balanced dataset. The ratio of minority data points might be higher if the sample size is increased, although this might not always be the case.

3.2 Resample the dataset

Resampling the dataset is one of the most frequently used method of all when it comes to issues regarding unbalanced datasets. A dataset is considered unbalanced when the majority class has much more data points than the minority class, or in other words, the minority class has drastically less data points than that of the majority class. Therefore, the intuitive way for balancing the dataset is to either increase the number of minority data points, called oversampling, or to decrease the number of data points in the majority class, termed undersampling. The spirit of resampling the dataset is to achieve a more balanced dataset through means of modifying the number of instances in classes.

3.2.1 Oversampling

Oversampling refers to the method of increasing the number of minority class instances. One of the frequently used approach is to make copies of existing instances.

3.2.2 Undersampling

Contrary to oversampling, undersampling makes changes to the majority class. By decreasing the amount of the majority class, a more balanced dataset can therefore be achieved. However, the main drawback of under-sampling is that potentially useful information contained in these ignored examples is neglected.[4]

3.3 Generate synthetic data

Generating synthetic data is a variation of oversampling, one of oversampling. In the previous section, oversampling is done in a more naive fashion, which is to simply make copies. Consider a method which utilizes not only the minority data points but also the relationship between the minority instances. SMOTE is the most popular method for generating synthetic data.

3.3.1 SMOTE

SMOTE is an algorithm for generating synthetic data. Suppose we have 10 minority class instances, and we want to create another 20.

First, randomly choose an instance from the minority class, suppose its called point χ.

Next, find k-nearest neighbors to the point chosen. Since we are creating another 20, 2 of the k neighbors are chosen. Take the difference between χ and one of the neighbors, and multiply the difference with a random number between 0 and 1.

Then add the difference to χ, creating a new point. Do the same thing for the other neighbor. Repeat the procedure for every data point within the original minority class.

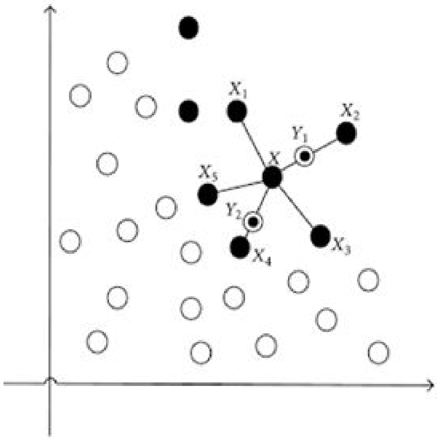


Diagram1. Graph representation of SMOTE[5]

.

Notes on Diagram 1: X is the chosen point, X1-X5 are the neighbors, Y1,Y2 are the points generated.

3.4 Change performance metric

The most commonly used metric for measuring a model's performance is its accuracy, which is the number of correctly classified instances. However, this is not the measurement suitable for dealing with unbalanced datasets. Imagine a dataset with 90% in Class A and 10% in Class B.The model created has a 90% accuracy, but all the Class B instances are all classified as Class A. Therefore, accuracy is not the best metric for measuring a model's performance. Other measurements, such as f1-score, recall , and precision are more appropriate for these kind of situations.

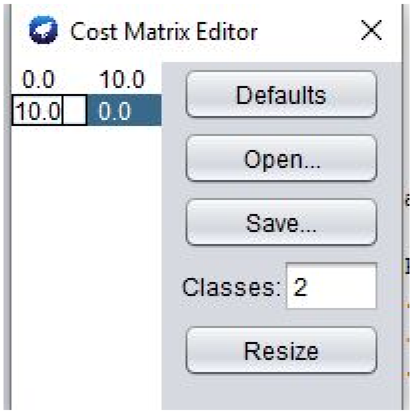
Precision: A measure of a classifiers exactness.

Recall: A measure of a classifiers completeness

F1 Score (or F-score): A weighted average of precision and recall.[6]

3.5 Utilize penalized models

Some models are designed as such that wrongly classified instances are to be penalized more severely. This helps enhance the chances of minority classes being recognized.



0.0	10.0
10.0	0.0

Classes: 2

Fig 5. Cost Matrix Editor in Weka

4. Algorithm performances on imbalanced datasets

It is agreed that imbalance datasets are difficult to deal with. Some algorithms perform better than others in imbalanced datasets. Decision trees often perform well on imbalanced datasets. The splitting rules that look at the class variable used in the creation of the trees, can force both classes to be addressed.[7]

There are also some other algorithms for classification that are more prone to imbalanced datasets. One of them is naive-bayes algorithm for classification. Therefore, the conventional naive-bayes method requires improvement if it is to be used for modeling imbalanced datasets.[8]

5. Experiment

We conducted an experiment on two datasets, the kaggle Credit Card Fraud Detection Dataset[9] and a Pima Diabetes dataset[10]. Our experiment is done on Weka, to compare the performance on the two dataset using two algorithms, Naive-Bayes and J48(C4.5) Decision Tree. 66% of the data were used as training set, the other 33% as test set.

5.1 Datasets

The Credit Card data set consists of several attributes, such as (1) number of seconds elapsed between this transaction and the first transaction (2)V1~V28 hidden attributes (3) amount of money transferred.

The Pima dataset consist of 8 attributes, such as (1) number of times pregnant, (2) plasma glucose concentration two hours in an oral glucose tolerance test (3) diastolic blood pressure (mm Hg), (4) triceps skin fold thickness (mm), (5) 2-hour serum insulin (mu U/ml), (6) body mass index (weight in kg/(height in m)^2), (7) diabetes pedigree function, and (8) age (years). This dataset is used as a not-so imbalanced imbalanced dataset.

5.2 Measurements

The measurements used are Correctly Classified Instances, and Weighted Average F-Measure.

5.3 Hypothesis

A significant difference between the algorithms used is expected. The Correctly Classified Instances metric may not tell us enough story to evaluate the performance of the algorithms. So we might observe a significant gap should occur in the F-Measure.

5.4 Result & Discussion

Correctly Classified Instances(%)

	Credit Card	Pima Diabetes
Naive Bayes	97.7074	77.0115
J48	99.9463	76.2452

Weighted Average F-Measure

	Credit Card	Pima Diabetes
Naive Bayes	0.987	0.769
J48	0.999	0.758

Turns out some of the results are quite unexpected. Decision tree as poorer performance overall when dealing with the Pima Indian dataset. However, in the Credit Card dataset, we can see an improvement on the Decision Tree algorithm. What we learned here is that in a very unbalanced dataset, decision tree might be a better choice of algorithm for doing classification. Naive Bayes has produced some unexpected surprisingly good result(quite far from what we have guessed), which deserves some more attention on why this happened.

Despite having a near perfect performance with the credit card dataset, we have a guess that overfitting might have occured. We still have to look deeper down into the dataset or the methods we used to find out why the performance was too good to be true. Another guess is that the tree shape has led to such a result, it could’ve been too deep..

6 Conclusion

From the experiment and research we've done this time, we've learned several lessons on data science worth noting. (1) Hypothesis don't always match the results of the experiment (2) When we ever encounter an imbalanced dataset, we should always be more careful, and think about the method that suits best with our dataset in hand.

7.References

[1]Pankaj Malhotra(2015, July 24), What is an imbalanced dataset?*,* *Quora* Retrieved March 15, 2019, from https://www.quora.com/What-is-an-imbalanced-dataset

[2]教育部統計處,*大專校院各校科系別學生數,* Retrieved March 15, 2019, from https://depart.moe.edu.tw/ed4500/News_Content.aspx?n=5A930C32CC6C3818&sms=91B3AAE8C6388B96&s=9D1CE6578E3592D7

[3] kaggle, *Credit Card Fraud Detection,* Retrieved March 16, 2019, from https://www.kaggle.com/mlg-ulb/creditcardfraud/version/3

[4]Vaishali Ganganwar(2012, April), An overview of classification algorithms for imbalanced datasets, *International Journal of Emerging Technology and Advanced Engineering,* Retrieved March 16, 2019, from https://pdfs.semanticscholar.org/239b/2210b3fbc1f4b8246437a88a668bf9a0d2c0.pdf

[5]Horace Tang(2017, August 28), 機器學習之陷阱 - Imbalance Class Classification, *Data Jungler,* Retrieved March 16, 2019, from http://datajungler.blogspot.com/2017/08/imbalance-class-classification.html

[6]Jason Brownlee(2015, August 19)*,*8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset*, Machine Learning Mystery,* Retrieved March 17, 2019, from https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset

[7]Jason Brownlee(2015, August 19)*,*8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset*, Machine Learning Mystery,* Retrieved March 17, 2019, from https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset

[8]Nur Maisarah Mohd Sobran, Arfah Ahmad, Zuwairie Ibrahim(2013, November 25), Proceeding of the International Conference on Artificial Intelligence in Computer Science and ICT(AICS 2013),25 -26 November 2013, Langkawi, MALAYSIA. (e-ISBN 978-967-11768-3-2). *Classification of imbalanced dataset using conventional naïve bayes classifier*, Paper presented at the International Conference on Artificial Intelligence in Computer Science and ICT Retrieved March 17, 2019, from https://pdfs.semanticscholar.org/2908/0eeec394d5958a2baf9d779491540130b6e7.pdf