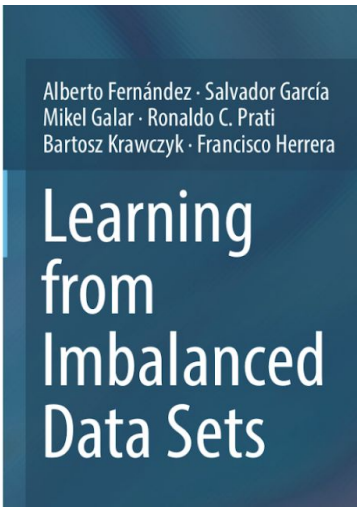
# IMBALANCED DATA E APLICAÇÕES

***<u>Quando os dados são extremamente desbalanceados, classificadores “normais” não funcionam!</u>***

Dados Desbalanceados: algumas áreas onde encontramos problemas reais.

- Predição de Acidentes automotivos (tráfego);
- Classificação de Dados Biomédicos;
- Detecção de Fraudes;
- Classificação de Imagens Médicas;
- Detecção de Faltas;
- Detecção de Anomalias;
- Diagnóstico Médico;
- Detecção de Vazamento de combustíveis;
- Muitas outras áreas.




Um conjunto de dados é considerado desbalanceado quando uma classe possui mais de 50% dos exemplos. Quando isto ocorre, muitos classificadores tendem a classificar de forma correta apenas os exemplos da classe majoritária (TOMAŠEV; MLADENIC´ , 2013). A degradação da classificação é mais intensa nos casos em que o nível de desbalanceamento é acima de 90% .(NAPIERALA; STEFANOWSKI,2016; PRATI; BATISTA; SILVA, 2015).

Métrica para dados não balanceados (IR): Imbalance Ratio (IR):é a razão entre o número de exemplos da classe minoritária pelo número total de exemplos do conjunto de dados do treinamento (NAPIERALA; STEFANOWSKI, 2016).

20 2 Foundations on Imbalanced Classification

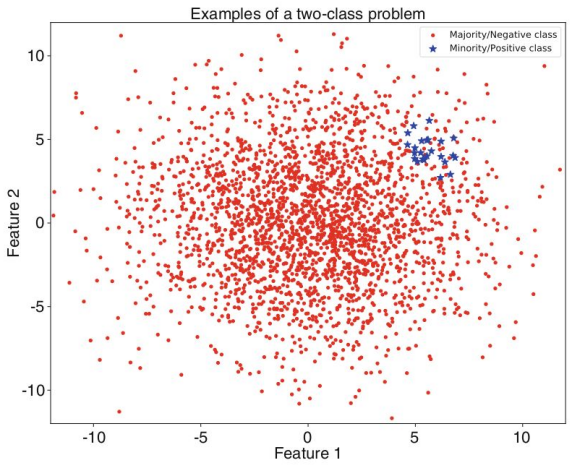



**Fig. 2.1** Example of a two-class imbalanced problem with ratio 1:100

Alguns estudos já realizados:

**Table 2.1** Applications of ML and DM where the class imbalance problem is present

| Year | Domain | Subcategory | Application | Data-level | Internal | Cost-sensitive | Ensemble | Reference |
|---|---|---|---|---|---|---|---|---|
| 1997 | Engineering | Satellite radar images | Detection of oil spills in satellite radar images | | × | | | [40] |
| 1997 | Engineering | Satellite radar images | Detection of oil spills in satellite radar images | × | | | | [41] |
| 1998 | Engineering | Satellite radar images | Detection of oil spills in satellite radar images | × | × | | | [42] |
| 2012 | Information technology | Software | Software defect prediction | × | | | × | [72] |
| 2013 | Bioinformatics | Protein identification | MicroRNA precursor classification | × | | | × | [45] |
| 2014 | Medicine | Quality control | Prediction of the post-operative life expectancy in lung cancer patients | | | × | × | [91] |
| 2014 | Bioinformatics | Protein identification | Five datasets that represent four different bioinformatics applications. These include miRNA identification, protein localization prediction, promoter identification from DNA sequences, kinase substrate prediction from protein phosphorylation profiling. | × | | | | [85] |
| 2014 | Information technology | Text mining | Text categorization | × | | | × | [82] |
| 2014 | Bioinformatics | Cell recognition | Mitotic cells recognition in Hep-2 images | × | | | × | [35] |
| 2014 | Medicine | Diagnosis | Lung nodule detection | × | | | × | [11] |
| 2014 | Information technology | Software | Software defect prediction | × | | × | × | [62] |
| 2014 | Security | Video surveillance | Face re-identification | × | | | × | [60] |
| 2014 | Information technology | Network analysis | Botnet traffic detection | × | | × | × | [30] |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2014 | Information technology | Network analysis | Network traffic classification | | | | × | [80] |
| 2015 | Medicine | Diagnosis | Breast cancer classification from thermogram images | | | × | × | [38] |
| 2015 | Information technology | Software | Software defect prediction | | | | × | [43] |
| 2015 | Bioinformatics | Protein identification | Contact map prediction in protein structure prediction | × | | | × | [78] |
| 2015 | Business management | Finance | Stock market prediction, credit card/loans approval, fraud detection | | × | | | [64] |
| 2015 | Medicine | Diagnosis | Automatic polyp detection | × | | | × | [6] |
| 2015 | Medicine | Quality control | Prediction of long stay patients in emergency department | | | | × | [4] |
| 2015 | Bioinformatics | Protein identification | Protein data classification | | | | × | [18] |
| 2015 | Medicine | Diagnosis | Diagnosis of diabetes mellitus | × | | | | [14] |
| 2016 | Business management | Customer relationship management | Customer churn prediction | × | | | | [3] |
| 2016 | Medicine | Diagnosis | Breast cancer malignancy classification | | | | × | [39] |
| 2016 | Medicine | Diagnosis | Bleeding detection in endoscopic video | × | | | × | [20] |
| 2016 | Education | High school | Early dropout detection | × | × | | | [53] |
| 2016 | Security | Video surveillance | Face re-identification | | × | | × | [68] |
| 2016 | Engineering | Semiconductors | Fault detection in semiconductors | × | | × | × | [44] |
| 2016 | Medicine | Diagnosis | Thyroid nodule classification | × | | | | [1] |
| 2016 | Medicine | Diagnosis | Breast cancer classification from Magnetic Resonance Images (MRIs) | | | | × | [52] |
| 2016 | Security | Biometric authentication | Multimodal biometric authentication | | | | × | [77] |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2017 | Engineering | Energy | Short-term voltage stability assessment | × | | × | | [90] |
| 2017 | Business management | Customer relationship management | Customer churn prediction | × | | × | × | [89] |
| 2017 | Information technology | Network analysis | Mobile malware detection | × | | × | | [15] |
| 2017 | Engineering | Semiconductors | Fault detection in semiconductors | × | | | × | [32] |
| 2017 | Medicine | Quality control | Prediction of the survival status of poly-trauma patients | × | | | | [65] |
| 2017 | Medicine | Prognosis | Prediction of bone fractures to prevent osteoporosis | × | | | | [5] |
| 2017 | Medicine | Diagnosis | Prediction of chronic kidney disease progression | × | | | | [16] |
| 2017 | Medicine | Prognosis | Donor-recipient matching prediction in liver transplantation | × | | | | [58] |
| 2017 | Security | Video surveillance | Still-to-video face recognition | | | × | | [8] |
| 2017 | Medicine | Diagnosis | Detection of microaneurysm | | | | × | [61] |
| 2018 | Engineering | Rotating machinery | Fault diagnosis in wind turbines | × | | × | × | [63] |
| 2018 | Information technology | Computer vision | Object recognition in images | | | | × | [87] |
| 2018 | Engineering | Rotating machinery | Fault diagnosis in wind turbines | | | × | | [73] |
| 2018 | Security | Video surveillance | Face re-identification | | | | × | [69] |
| 2018 | Engineering | Rotating machinery | Fault diagnosis in wind turbines | | | | × | [83] |

Separando apenas as áreas da saúde e correlatas temos as seguintes aplicações:

1. Classificação de MicroRNA;
2. Predição da expectativa de vida de pacientes operados de câncer de pulmão;
3. Detecção de nódulos no pulmão;
4. Classificação de câncer de mama através de imagens;
5. Detecção automática de pólipos;
6. Predição do tempo de permanência de pacientes na emergência;

7. Diagnóstico de Diabetes;
8. Classificação de malignidade de câncer de Mama;
9. Detecção de sangramento em imagens de endoscopia;
10. Classificação de nódulos na tireoide;
11. Classificação de cancer de mama através de imagens de ressonancia magnética;
12. Descoberta de compatibilidade entre doador e receptor de fígado para transplante;
13. Detecção de microaneurisma;