Dataset balancing and noise reduction

Rafael Baeta

Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

2 de março de 2017

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

UFMG

Introduction

- Data in the real world is dirty!
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names

Introduction

- Many problems in data occurs due to:
  - Data is not always available
  - Equipment malfunction
  - Data not entered due to misunderstanding
  - Certain data may not be considered important at the time of entry

Quality decisions must be based on quality data!

Introduction

- And these problems can lead to:
  - Unbalanced datasets
    - Some classes have a much smaller number of examples than other classes
  - Noise data
    - Some data can be similar and belong to different classes.
    - Data attributes can be corrupted or missing.

Introduction

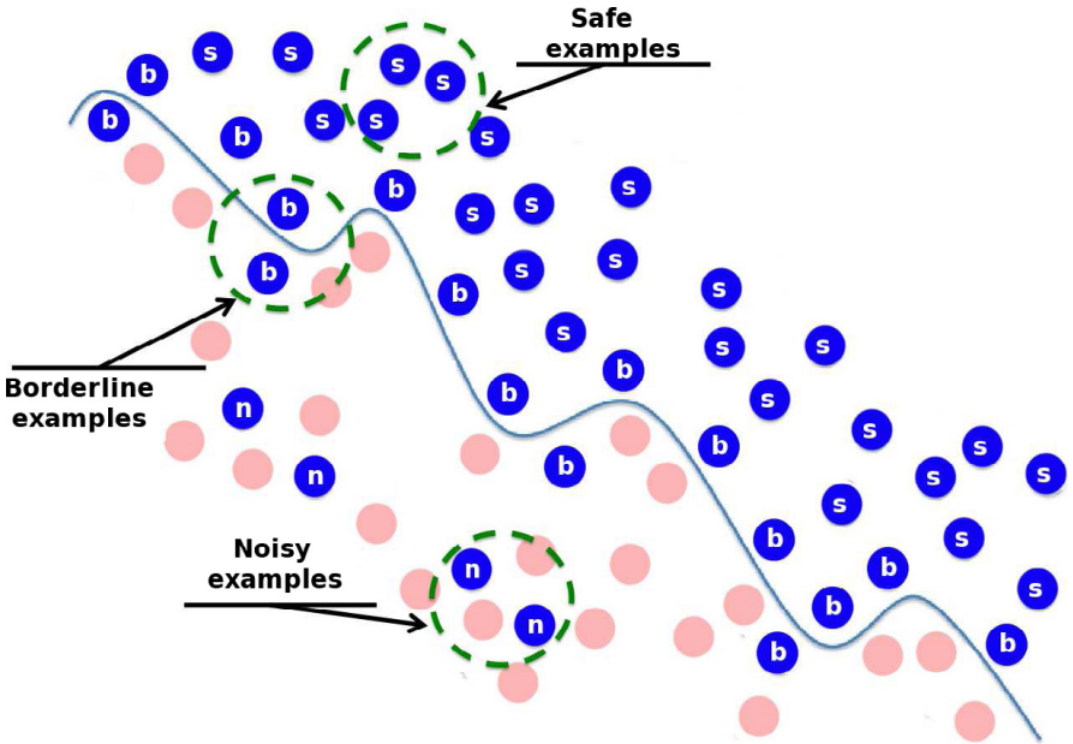


Figura : Unbalanced and noisy data

Motivation

- Datasets with high unbalanced data can lead to privilege some classes
- Some data can be wrong and can penalize the process of learning
- Some redudant data can be removed raising the training speed.
- The classifier becomes more general.

Dataset balancing

What to do if get more data for minority classes is not possible?

Dataset balancing

What to do if get more data for minority classes is not possible?

- Undersampling
  - Reduction of data that belongs to majority classes. How to select data to be removed?

Dataset balancing

What to do if get more data for minority classes is not possible?

- Undersampling
  - Reduction of data that belongs to majority classes. How to select data to be removed?
- Oversampling
  - Generate more data for the minority classes. How to generate this data?

Dataset balancing

What to do if get more data for minority classes is not possible?

- Undersampling
  - Reduction of data that belongs to majority classes. How to select data to be removed?
- Oversampling
  - Generate more data for the minority classes. How to generate this data?
- Weighting
  - Assigning larger misclassifying cost for minority class samples

Dataset balancing

What to do if get more data for minority classes is not possible?

- Undersampling
  - Reduction of data that belongs to majority classes. How to select data to be removed?
- Oversampling
  - Generate more data for the minority classes. How to generate this data?
- Weighting
  - Assigning larger misclassifying cost for minority class samples
- Ensemble methods
  - A combination of a variety of base classifiers

Undersampling

- An undersampling technique removes samples from the majority class
- It can remove relevant samples.
- A commom technique is random undersampling.
- Another techniques: Neighboorhood Cleaning Rule (NCR), Condensend Nearest Neighbour (CNN), One-Sided-Selection (OSS),[1] etc.

[1] Miroslav Kubat, Stan Matwin et al. “Addressing the curse of imbalanced training sets: one-sided selection”. Em: *ICML*. vol. 97. Nashville, USA. 1997, pp. 179–186.

Undersampling

- This approach[2] define k clusters.
- Calculate the number of majority examples extracted from each cluster based on a proportion between the majority and minority classes.
- Randomly select majority class samples within each cluster.
- The majority samples can be selected using heuristics of distances.

[2] Show-Jane Yen e Yue-Shi Lee. "Cluster-based under-sampling approaches for imbalanced data distributions". Em: *Expert Systems with Applications* 36.3 (2009), pp. 5718–5727.

Undersampling

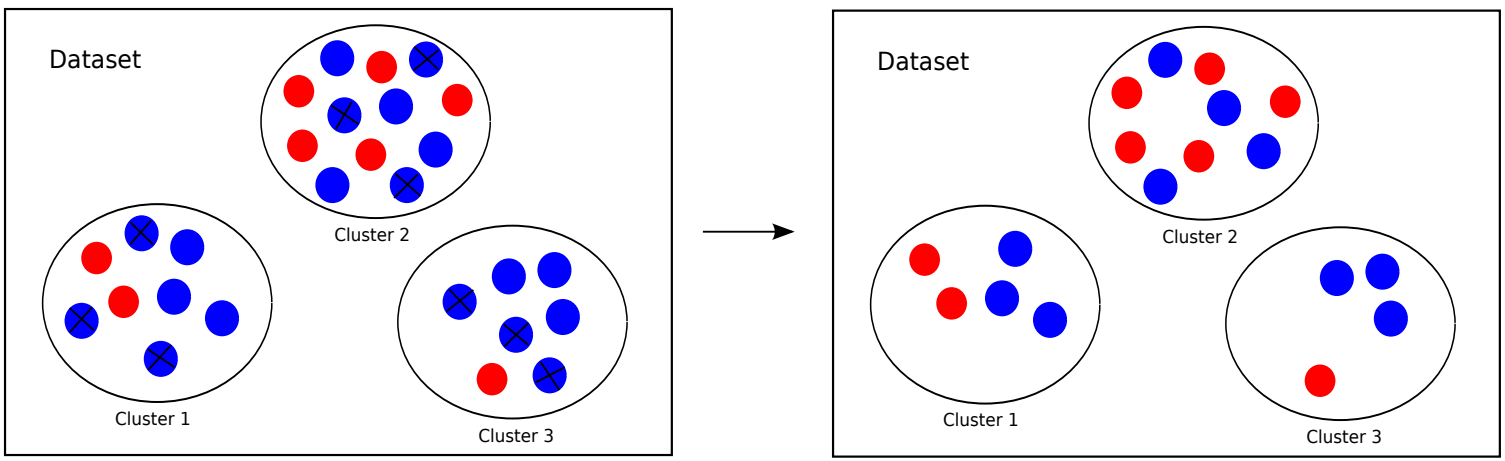


Figura : Cluster undersampling

Oversampling

- SMOTE[3] is a popular oversampling technique for binary problems
- Minority class is over-sampled by creating "synthetic examples".
- Extra training data is created by operating in the "feature space".

[3] Nitesh V. Chawla et al. “SMOTE: synthetic minority over-sampling technique”. Em: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

Oversampling

- Synthetic samples are generating in the following way:
  - Take the difference between the feature vector (sample) under consideration and its neighbor.
  - Multiply this difference by a random number between 0 and 1 and add it to the feature vector under consideration
  - This causes the selection of random points along the line segment between specific features.

Oversampling

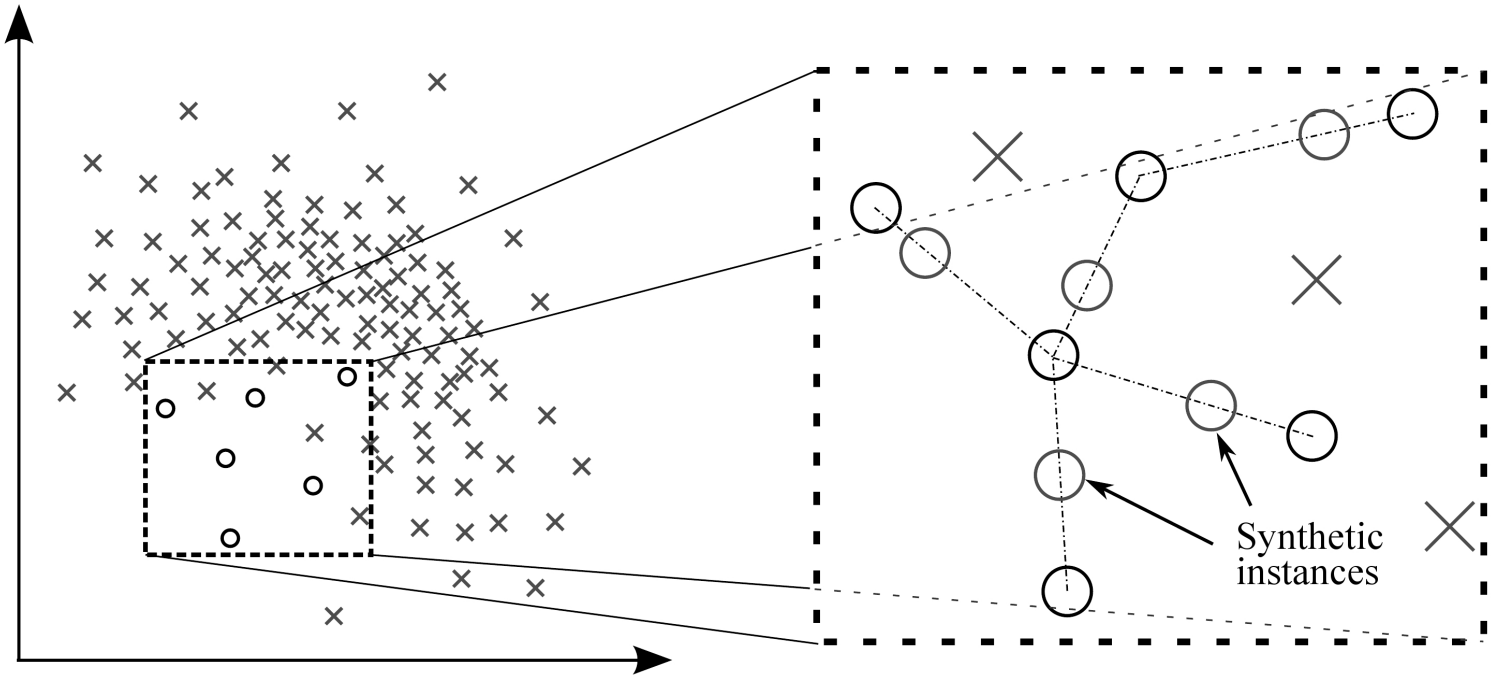


Figura : Smote example

Oversampling

- In[4] is created a approach to deal with multiclass inbalance
- This approach uses the SMOTE algorithm to generate syntethic samples
- The synthetic samples are created based in some configurations that can be "outlier", "rare", "safe", "bordeline"

[4] José A Sáez, Bartosz Krawczyk e Michał Woźniak. “Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets”. Em: *Pattern Recognition* 57 (2016), pp. 164–178.

Oversampling

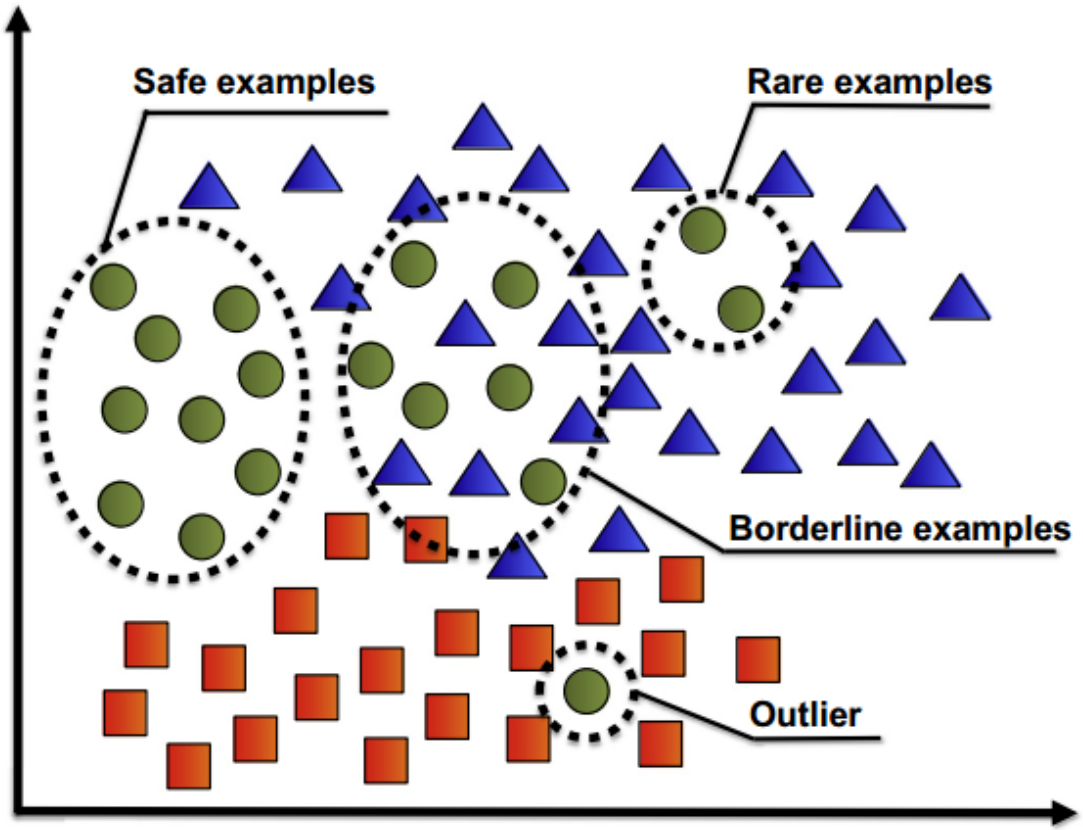


Figura : Smote example

Oversampling

- The samples can be classified as "rare", "outlier", "safe", "bordeline"depending on the number of examples (cn(e)) that share the class label.
  - Safe examples: $cn(e) >= 4$
  - Borderline examples: $2 >= cn(e) >= 3$
  - Rare examples: $cn(e) = 1$
  - Outliers examples: $cn(e) = 0$

Oversampling

- The synthetic examples are created as follow:
  - A random example x of a class and type are iterativily selected
  - Then, a random example y within the k=5 nearest neighbors is selected
  - With these two examples, x and y, the new synthetic example is created by interpolation as SMOTE does.

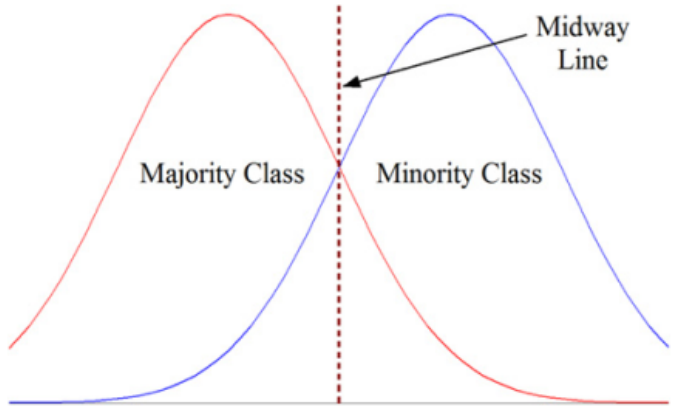
Weighting

- Focus on modifying existing classification algorithms to stregthen the learning of minority class.[5]
- Most of algorithms in this family are based on SVM and Neural Network[6]

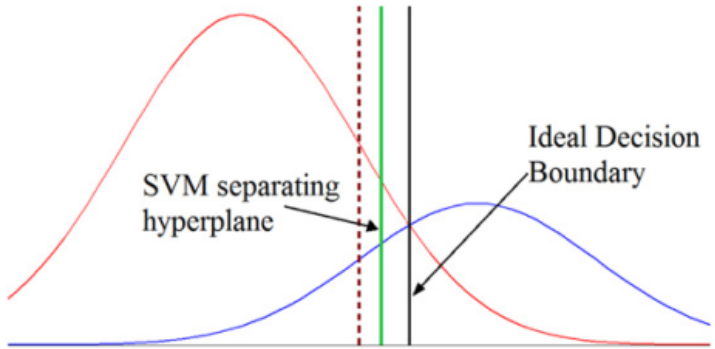
[5] Shounak Datta e Swagatam Das. "Near-Bayesian Support Vector Machines for imbalanced data classification with equal or unequal misclassification costs". Em: *Neural Networks* 70 (2015), pp. 39–52.

[6] Li Yijing et al. "Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data". Em: *Knowledge-Based Systems* 94 (2016), pp. 88–104.

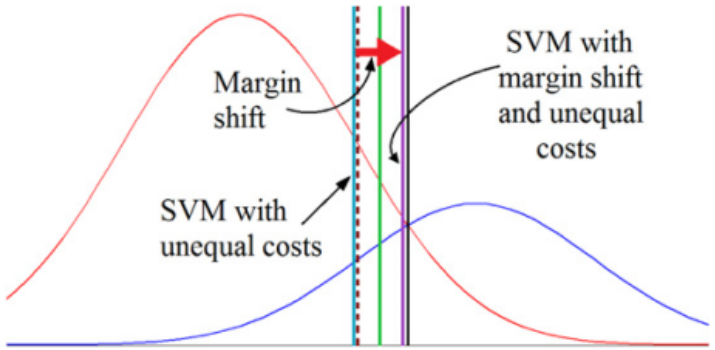
Weighting



(a) Class-conditional probability distributions for the majority class and minority class.



(b) Posterior probability distributions for the two classes. The separating hyperplane obtained by SVM is closer to the minority class, but does not coincide with the ideal decision boundary.



(c) Adding unequal costs for the two classes moves the hyperplane close to the midway line, this hyperplane can be shifted towards the ideal decision boundary using proper biasing.

Figura : SVM margin shift with unequal costs

Weighting

- Traditional SVM - $\min_{w,b} \frac{1}{2}\|w\|^2$
- Soft SVM - $\min_{w,\xi,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi$
- Soft SVM with unequal regularization - $\min_{w,\xi,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} f(X_i)\xi$
  - $f(X_i) = \begin{cases} \frac{1}{p_+}, & \text{if } x_i \in I_+ \\ \frac{1}{p_-}, & \text{if } x_i \in I_- \end{cases}$

Ensemble

- This methodology[7] is composed of the following steps:
  - The imbalanced training set is submeted to a feature selection

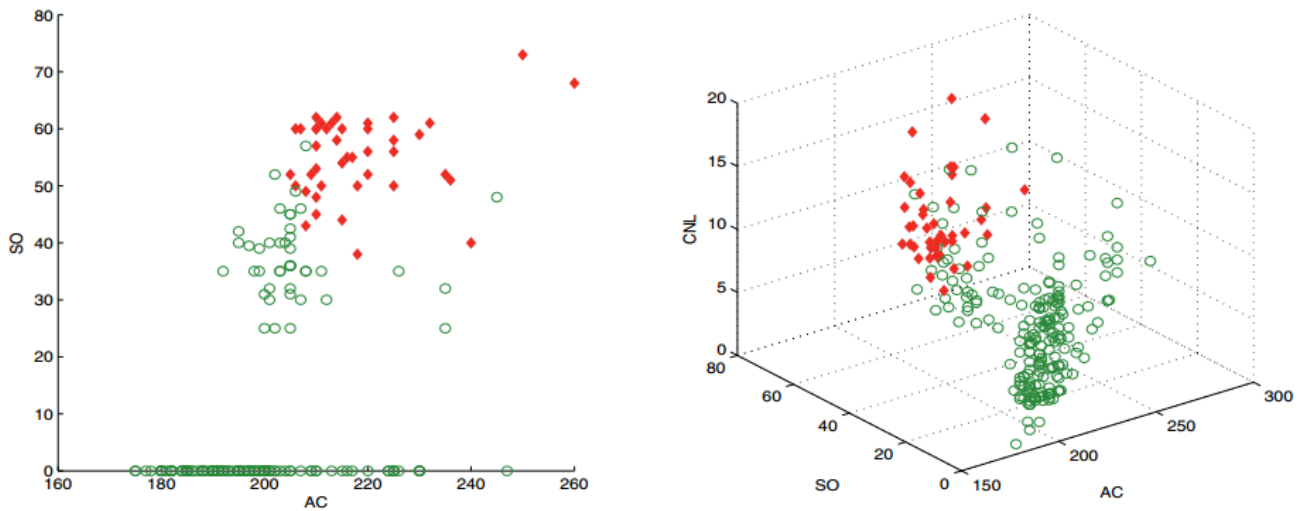


Figura : Feature selection

[7] Yijing et al., “Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data”.

Ensemble

- Then the training set is divided in k subtraining sets
- Each training set is give to train a classifier.
- The answer of the classifiers are submeted to a ensemble rule.

Ensemble

- This approach test 5 ensemble rules: max, min, product, majority vote and sum.
- Max - $\underset{1\leq j\leq n}{\text{argmax}}\{R_{C_1}, R_{C_2}, ..., R_{C_m}\}, R_{C_j} = \max\limits_{1\leq i\leq k} AUCarea_i.p_{ij}$
- Min - $\underset{1\leq j\leq n}{\text{argmax}}\{R_{C_1}, R_{C_2}, ..., R_{C_m}\}, R_{C_j} = \min\limits_{1\leq i\leq k} AUCarea_i.p_{ij}$
- Product - $\underset{1\leq j\leq n}{\text{argmax}}\{R_{C_1}, R_{C_2}, ..., R_{C_m}\}, R_{C_j} = \prod_{i=1}^{K} AUCarea_i.p_{ij}$
- Majority vote - $\underset{1\leq j\leq n}{\text{argmax}}\{R_{C_1}, R_{C_2}, ..., R_{C_m}\}, R_{C_j} = \underset{1\leq i\leq k}{\text{count}}\left(\underset{c_1,c_2,...,c_m}{\text{argmax}} \text{AUCarea}_\text{i}.\,\text{p}_\text{ij}\right)$
- Sum - $\underset{1\leq j\leq n}{\text{argmax}}\{R_{C_1}, R_{C_2}, ..., R_{C_m}\}, R_{C_j} = \underset{1\leq i\leq k}{\text{sum}} AUCarea_i.p_{ij}$

Ensemble

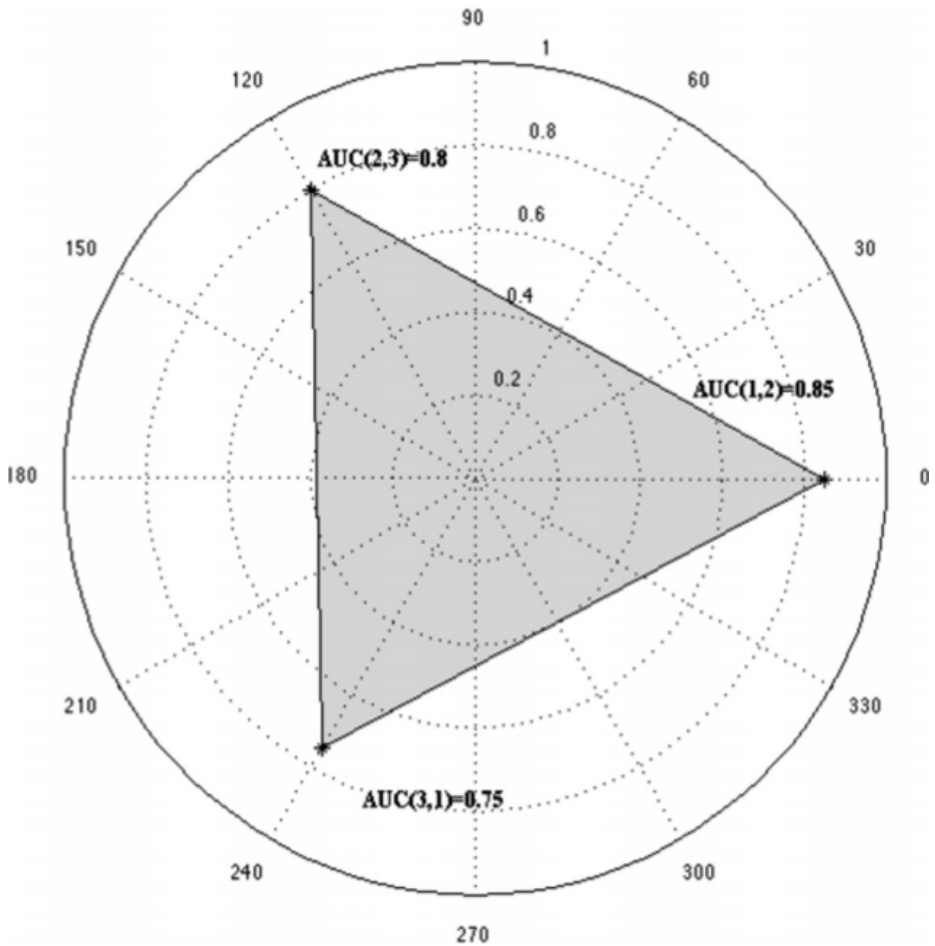


Figura : AUCarea metric

Ensemble

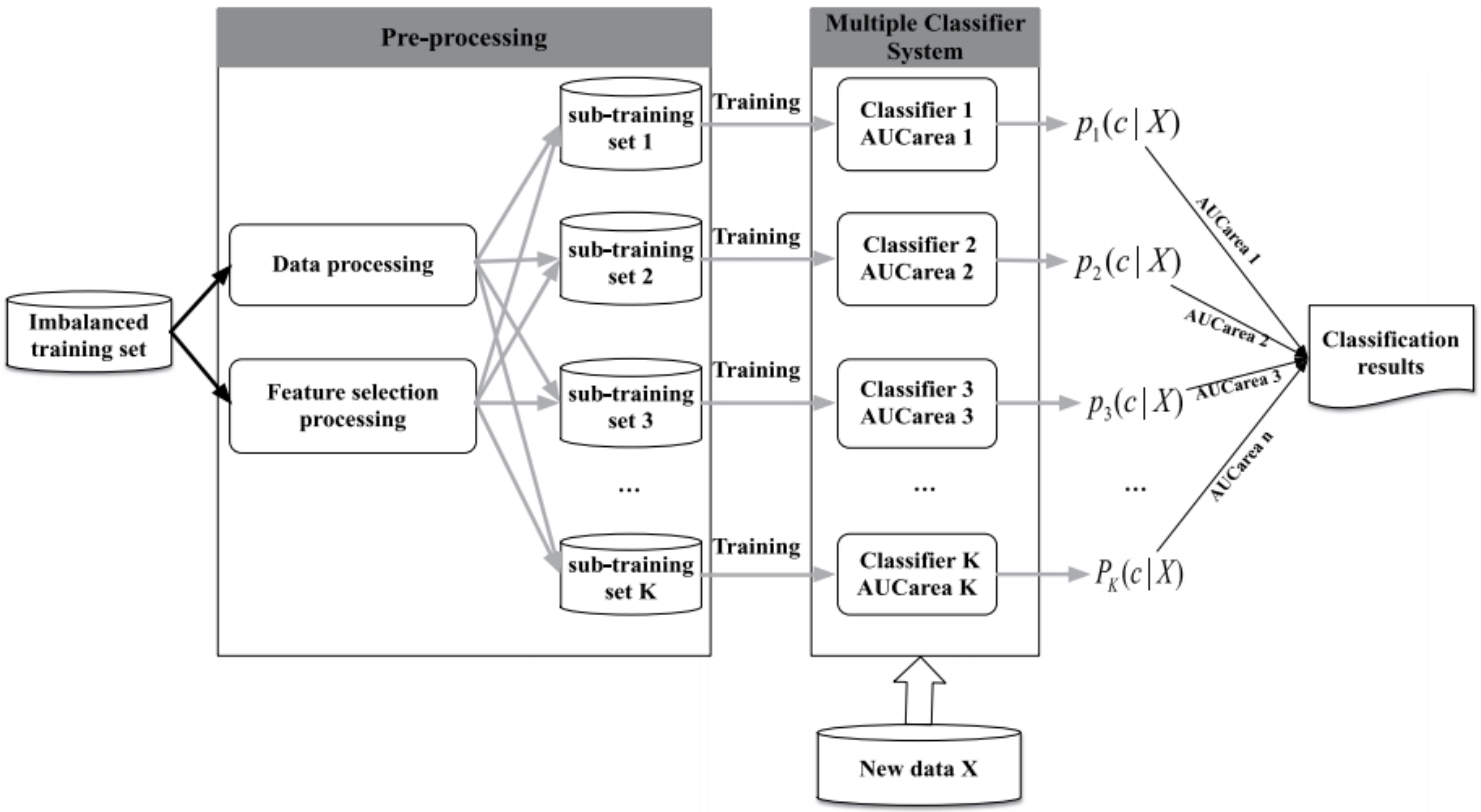


Figura : Architecture of ensemble

Ensemble

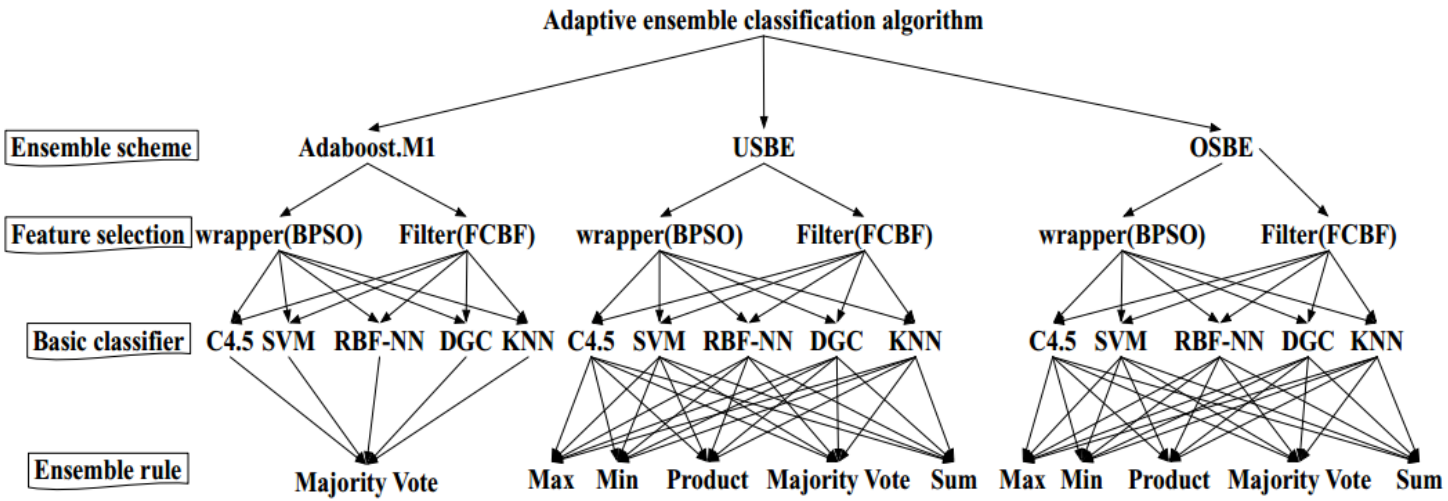


Figura : Adapting path of AMCS

Conclusion

- Dataset balancing is essential to avoid overfitting in classifiers.
- Remove useless samples
- Make classifiers more general

References I

Batista, Gustavo EAPA, Ronaldo C Prati e Maria Carolina Monard. “A study of the behavior of several methods for balancing machine learning training data”. Em: *ACM Sigkdd Explorations Newsletter* 6.1 (2004), pp. 20–29.

Cateni, Silvia, Valentina Colla e Marco Vannucci. “A method for resampling imbalanced datasets in binary classification tasks for real-world problems”. Em: *Neurocomputing* 135 (2014), pp. 32–41.

Chawla, Nitesh V. et al. “SMOTE: synthetic minority over-sampling technique”. Em: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

Datta, Shounak e Swagatam Das. “Near-Bayesian Support Vector Machines for imbalanced data classification with equal or unequal misclassification costs”. Em: *Neural Networks* 70 (2015), pp. 39–52.

Kubat, Miroslav, Stan Matwin et al. “Addressing the curse of imbalanced training sets: one-sided selection”. Em: *ICML*. Vol. 97. Nashville, USA. 1997, pp. 179–186.

References II

Rahman, M Mostafizur e DN Davis. “Addressing the class imbalance problem in medical datasets”. Em: *International Journal of Machine Learning and Computing* 3.2 (2013), p. 224.

Sáez, José A, Bartosz Krawczyk e Michał Woźniak. “Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets”. Em: *Pattern Recognition* 57 (2016), pp. 164–178.

Yen, Show-Jane e Yue-Shi Lee. “Cluster-based under-sampling approaches for imbalanced data distributions”. Em: *Expert Systems with Applications* 36.3 (2009), pp. 5718–5727.

Yijing, Li et al. “Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data”. Em: *Knowledge-Based Systems* 94 (2016), pp. 88–104.