Could Functional Dependencies Help to Identify Balanced Classification Datasets ?

Marie Le Guilly, Jean-Marc Petit and Marian Scuturici

Motivations

Machine learning and databases:

- More and more associated together...
- ... but still quite different domains when taken individually.
- It could be interesting to extend their intersection !

Table of contents

Imbalanced datasets and classification

Imbalanced datasets

Imbalanced dataset are a recurring problem in machine learning.

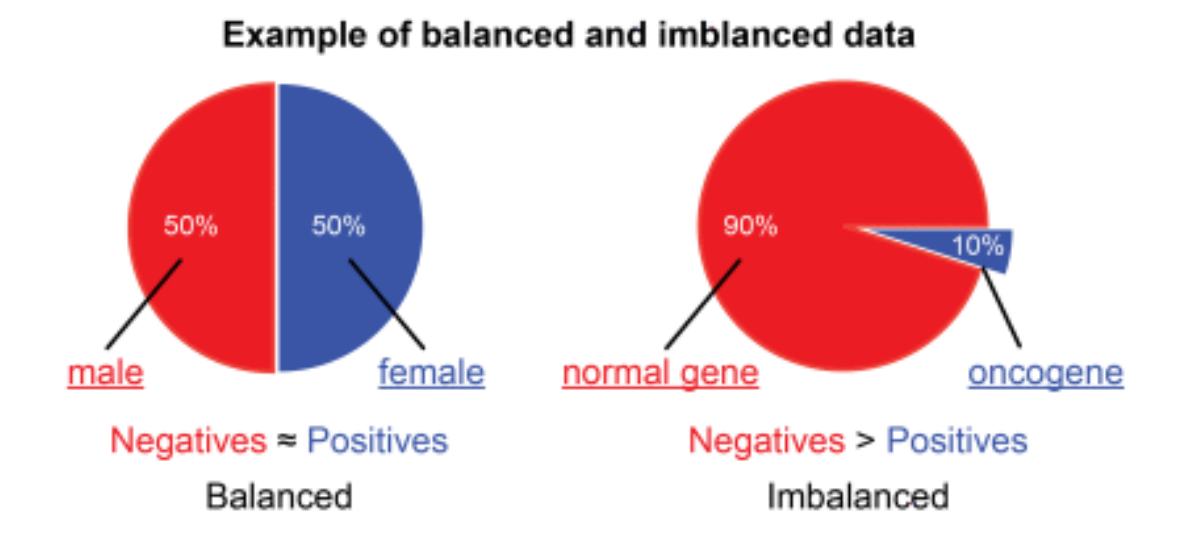


Figure 1: Balanced versus imbalanced dataset

Classifier (almost) always predicts the largest class...

... while the smaller is more interesting.

Existing solutions

Solutions have been proposed to tackle the problem of imbalanced datasets: for example **sampling** methods.

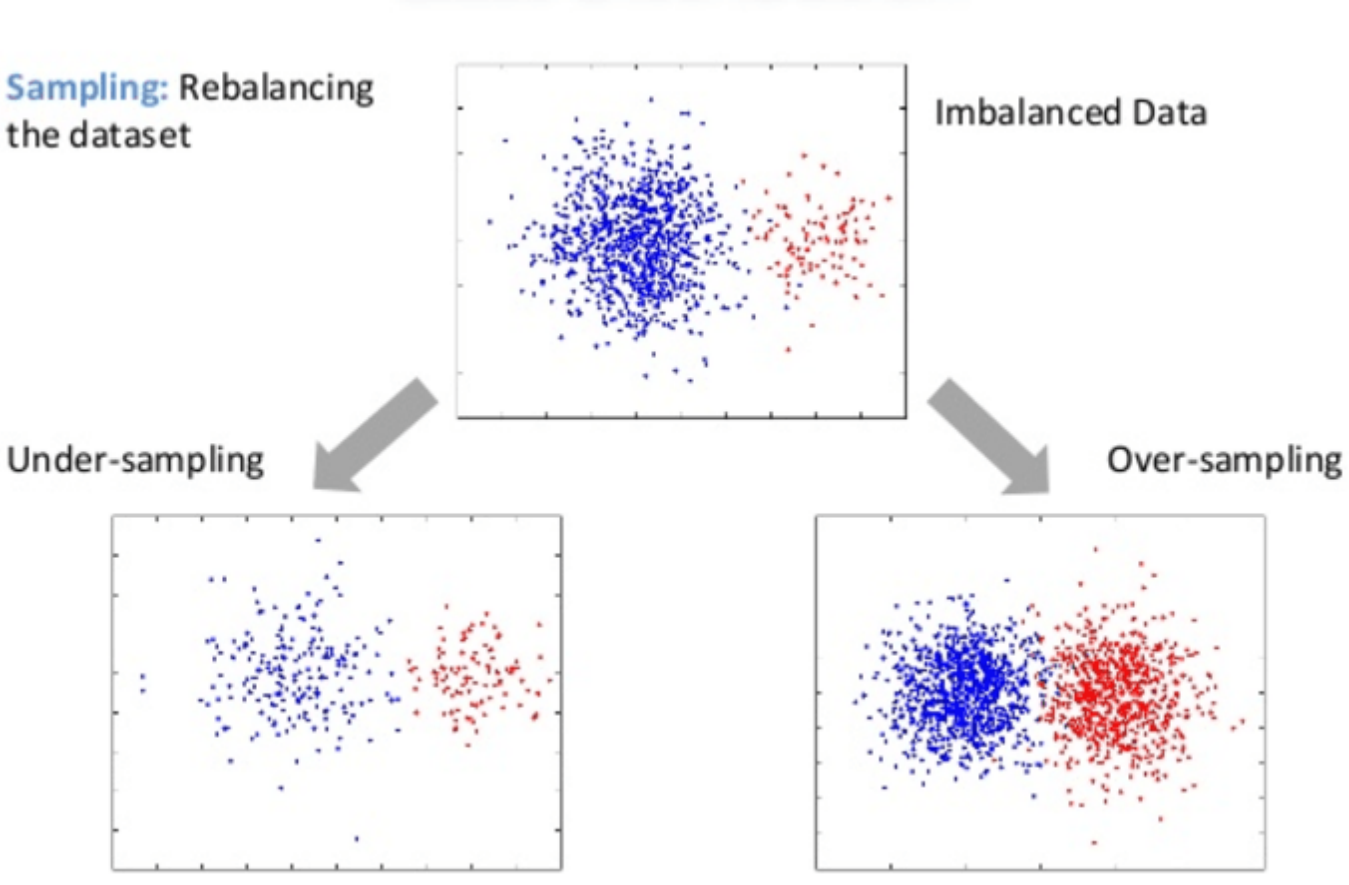


Figure 2: The two main sampling techniques for imbalanced datasets

Formalization

Let's consider the binary classification of a dataset r:

- Tuples labeled as negative are in $r^- \subseteq r$
- All other are labeled as positive and in $r \setminus r^-$
- r is imbalanced i.e. $|r^-| \ll |r \setminus r^-|$

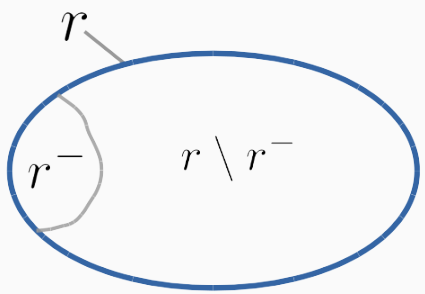


Figure 3: Illustration of the considered binary classification problem

Undersampling

Undersampling approach to the problem:

1. Find a subset $r^+ \subseteq r \setminus r^-$
   such that $|r^+|$ is similar to $|r^-|$
2. And then classify between r^+ and r^- !

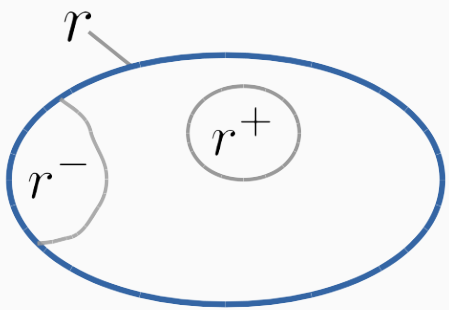


Figure 4: Illustration of the undersampling solution

The choice of samples is often based on statistical techniques (see [KKP+06]).

Imbalanced datasets in databases

In databases, some specific situations present an imbalanced dataset problem: in [Cumin et al. 2017], a decision tree is used to reformulate an SQL query.

- All tuples from the SQL query are considered as one class
- The reservoir of available tuples for the other class is enormous !

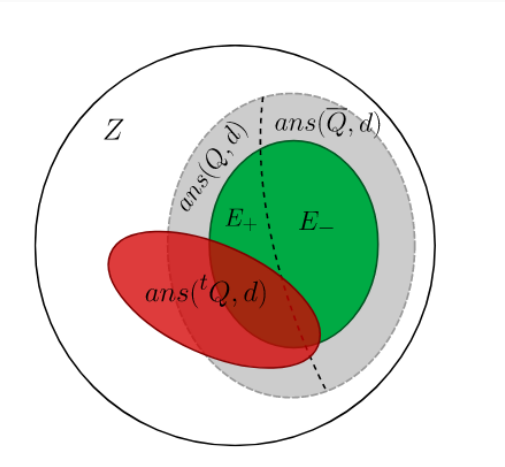


Figure 5: Imbalanced dataset problem in query reformulation

Proposed solution: compute a negation of the SQL query with a similar answer set's size.

Functional Dependencies

It could be interesting to take a completely different look at this problem.

Functional dependencies are powerful objects, representing constraints and implications in data.

They are really important for relational databases, for example for the normalization of database's schema...

... But not really considered in machine learning !

Proposal: select samples with a completely different approach, by considering the **functional dependencies** of sets r^+ and r^-, and their interactions.

Preliminaries

Functional Dependencies

Syntax

A FD on a schema $\mathcal{R}$ is a declaration of the form:

$$r : X \to Y, \text{ where } X, Y \subseteq \mathcal{R}.$$

A FD can then be satisfied or not by a relation:

Semantic

A FD $r : X \to Y$ is satisfied by a relation r on $\mathcal{R}$, i.e. $r \models X \to Y$, if and only if $\forall t_1, t_2 \in r$:

$$\text{if } t_1[X] = t_2[X] \text{ then } t_1[Y] = t_2[Y].$$

Example:

r	A	B	C
	2	3	5
	2	3	8
	4	1	5

We have :

$r \models A \to B$

but $r \not\models A \to C$

Closure of attributes

Definition 1

Let $X \subseteq \mathcal{R}$ and F a a set of FDs on $\mathcal{R}$. The closure of X w.r.t F, X_F^+ is defined as :

$$X_F^+ = \{A \in R \mid F \models X \rightarrow A\}$$

where $\models$ means "logical implication"

Definition 2

X is closed w.r.t F if $X_F^+ = X$.

Definition 3

The closure system $CL(F)$ is the set of closed sets of F :
$CL(F) = \{X | X = X_F^+\}$

This definitions generalize to any relation r, F becoming implicit.

From a relation it is possible to get a closure system, and vice-versa.

On the distance of databases

Notion of Distances in databases

Distance in databases: not many resources on the subject !

- Number of differences between two databases [MFL06]: but not symmetric and more for comparing between an original database and its modified version.
- In terms of functional dependencies : let's see how ! [KKS10].

Distance of databases

The **distance** of databases is then defined as follows:

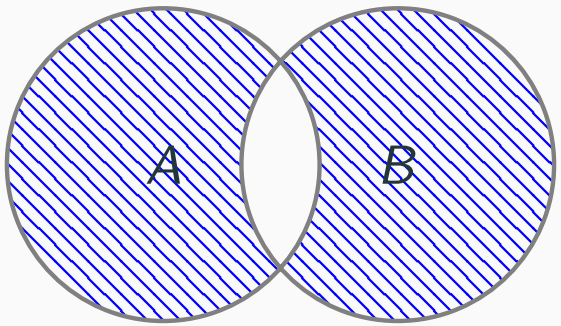
Definition 4

The distance between two instances r_1 and r_2 of the schema R, with respective sets of FDs F_1 and F_2 is:

$$d(r, r') = |CL(F_1) \triangle CL(F_2)|$$

where $A \triangle B$ denotes the symmetric difference of the two sets, i.e:

$$A \triangle B = A \setminus B \cup B \setminus A.$$



Example

r_1	A	B	C
	0	0	0
	3	0	3
	1	1	1
	0	0	4
	0	2	2

r_2	A	B	C
	0	0	0
	4	0	0
	2	2	0
	0	0	5
	0	3	0

$CL(r_1) = \{ABC, AB, A, B, \emptyset\}$ $CL(r_2) = \{ABC, AB, AC, BC, A, B, C, \emptyset\}$

So we obtain:

$$d(r_1, r_2) = |\{AC, BC, C\}| = 3$$

Maximum Distance

The distance between two databases has an upper bound:

Property 1 (Katona et al. 2010)
Let $\mathcal{R} = n$. Then, for any two instance r_1 and r_2 of schema $\mathcal{R}$:

$$d(r_1, r_2) \leq 2^n - 1$$

This is the maximum number of closed elements, minus the top one which is in any closed set.

Intuition of distance of databases

This definition of distance can be a bit... surprising !

Let's try to get the intuition of what it means with a basic example:

- $CL_1 = \{ABC, AB, AC, A\}$
- $CL_2 = \{ABC, BC, B, C, \emptyset\}$
- for two relation r_1 and r_2 with respective closure systems CL_1 and CL_2, we get :

$$d(r_1, r_2) = |\{AB, AC, A, BC, B, C, \emptyset\}| = 2^3 - 1 = 7$$

- $r_1 \models \{BC \to A\}$
- $r_2 \models \{A \to BC\}$

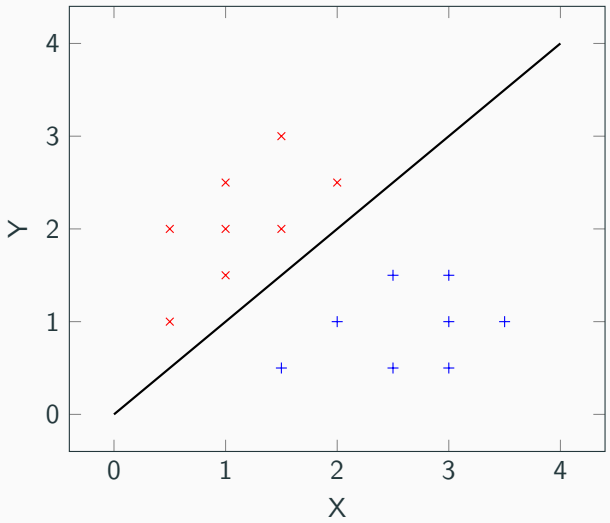
Distant relation satisfy "opposite" functional dependencies !

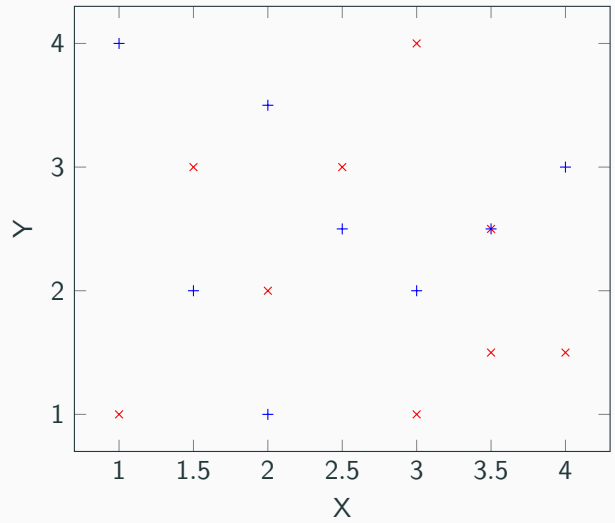
Back to classification

Distance and classification

Conjecture: it is easier to classify between distant sets.

Let m be a classification model, and $score_m(r^+, r^-)$ be the score of this classifier when discriminating between tuples from two relation r^+ and r^-. Then, when $d(r^+, r^-)$ increases, so does $score_m(r^+, r^-)$, and conversely.





Distant set and imbalanced classification

Considering that it is easier to classify between distant datasets (in terms of DF): **use it to construct a balanced dataset !**

If we come back to the initial problem: considering r_+, find r^- in $r \setminus r^+$ such that:

- $|r^+| \approx |r^-|$
- ... and $d(r^+, r^-)$ is maximized !

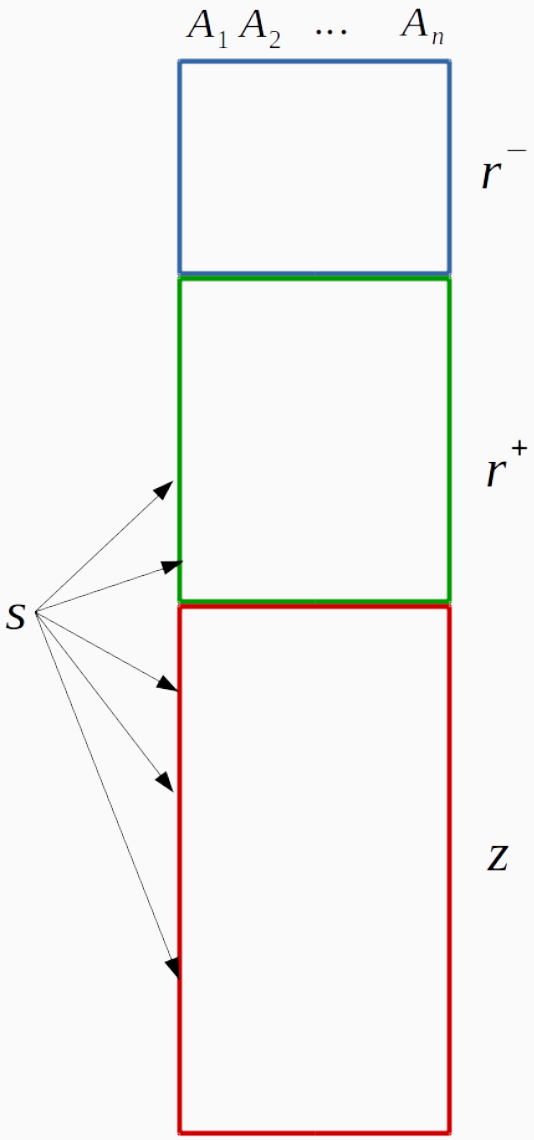
The sampling of the majority class would be based on this notion of distance in databases.

Problem

We first want to answer the following existential question:

Is it possible to find instances such that the conjecture is true ?

Problem statement



Is it possible to find two relations r^- and r^+ such that $d(r^+, r^-)$ is as large as possible, such that :

- $d(r^+, r^-) \geq d(s, r^-)$
- $score_m(r^+, r^-) \geq score_m(s, r^-)$?

Where s is a random sample of tuples from $z \cup r^+$ with $|s| \simeq |r^-|$.

z is the set of all possible tuples on $(adom(r^+))^n$

Proposition

To answer our question, we have to:

Step 1: For a given schema of size n, create two closure systems CF^+ and CF^-, such that $|CF^+ \triangle CF^-| = 2^n - 1$ (maximum), and $|CF^+| \approx |CF^-|$.

Step 2: Derive two relations r^+ and r^- with respective closure systems CF^+ and CF^-

Step 3: Apply a classification model to discriminate between tuples from r^+ and r^-.

Step 4: Create a relation s, with random values on the same active domain as the one of r^+, such that $|s| \approx |r^-|$.

Step 5: Apply classification model on r^- versus s.

Element of solution

Our solutions requires the study of several sub-problems.

- The generation of two closure systems is a difficult combinatorial problem: not treated in this presentation.
- Data generation from closure systems relies on Armstrong relations.
- Synthetic data generation can rely on several strategy: see implementation.
- Several classification algorithms can be tested, see experimentations.

Implementation and Experimentations

Data generation: r+ and r-

Example on a schema of size 4:

r^-	A	B	C	D
	0	0	0	0
	0	0	0	1
	2	0	0	0
	0	3	0	3
	4	0	0	4
	5	5	0	0
	6	6	0	6

r^+	A	B	C	D
	7	7	7	7
	7	7	8	7
	7	9	7	7
	7	7	10	10
	7	11	11	7
	12	7	12	7
	7	13	13	13
	14	7	14	14
	15	15	15	7
	16	16	16	16

$$CL^- = \{ABCD, ABC, BCD, AC, BC, CD, C\}$$

$$CL^+ = \{ABCD, ABD, ACD, AB, AD, BD, A, B, D, \emptyset\}$$

$$d(r^+, r^-) = 2^4 - 1 = 15$$

Data generation: Z and s

Z is generated from the vast reservoir of possible tuples: with values on the active domain of r^- and r^+.

s is then a random sample from Z.

s	A	B	C	D
	7	3	0	8
	4	8	2	14
	11	10	1	11
	11	13	1	7
	0	6	6	3
	11	1	2	8
	6	11	15	2

$ADOM(Z) = ADOM(r^+ \cup r^-)$

Test conditions

Test conditions:

- Schema of size 12
- Data generated from closure systems, with a randomized order on integer values.
- Test on 10 classifiers

Set	r^-	r^+	s
Training test size	1361	1916	1361
Testing set size	341	479	341
Total	1702	2395	1702

Results: simple classification

The first experimentation aims at studying classification scores in a "classic" setting, with one dataset with distant classes.

We therefore train to models:

- M_1 on r^- and r^+
- M_2 on r^- and s

They are both evaluated on the testing sets of their respective datasets.

Classifier	r_- **vs** r^+	r^- **vs** s
Nearest Neighbors	0.91	0.74
Decision Tree	0.99	0.96
Random Forest	1.0	0.99
AdaBoost	0.99	0.99
Neural Net	0.76	0.67
Naive Bayes	1.0	0.75
QDA	0.87	0.81
Gaussian Process	0.48	0.33
RBF SVM	0.77	0.70

Results: imbalanced conditions

In the imbalanced problem, the model is trained on balanced classes, but the purpose is to evaluate on the general dataset with imbalanced classes.

Now, both M_1 and M_2 are evaluated on the testing set of r^- and Z.

Classifier	r_- **vs** r^+	r^- **vs** s
Nearest Neighbors	0.91	0.74
Decision Tree	0.99	0.96
Random Forest	1.0	0.99
AdaBoost	0.99	0.99
Neural Net	0.76	0.67
Naive Bayes	1.0	0.75
QDA	0.87	0.81
Gaussian Process	0.48	0.33
RBF SVM	0.77	0.70
Linear SVM	0.67	0.47

Conclusion

Conclusion... and many questions

This in an explorative study:

- Experimental results tend to contradict the initial intuition
- But this is one possible solution for the initial motivation with SQL queries

Functional dependencies do capture global constraints on data, and are an elegant solution when no other information is available.

In addition, our approach could be used to build synthetic datasets that are hard to classify.

Future work could focus more on algorithms detecting FDs in dataset, to reinforce value-based approaches.

Thank you !

References I


Catriel Beeri, Martin Dowd, Ronald Fagin, and Richard Statman.
On the structure of armstrong relations for functional dependencies.
Journal of the ACM (JACM), 31(1):30–46, 1984.


Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al.
Handling imbalanced datasets: A review.
GESTS International Transactions on Computer Science and Engineering, 30(1):25–36, 2006.


Gyula OH Katona, Anita Keszler, and Attila Sali.
On the distance of databases.
In *International Symposium on Foundations of Information and Knowledge Systems*, pages 76–93. Springer, 2010.

References II


Heiko Müller, Johann-Christoph Freytag, and Ulf Leser.
Describing differences between databases.
In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 612–621. ACM, 2006.