Dataset Characterization for CIVEMSA 2018 Paper

Gerardo Felix[a], Gonzalo Nápoles[b], Rafael Falcon[c],
Rafael Bello[a], Koen Vanhoof[b]

[a] *Department of Computer Science, Universidad Central de Las Villas, Cuba*
[b] *Faculty of Business Economics, Hasselt Universiteit, Belgium*
[c] *School of Electrical Engineering and Computer Science, University of Ottawa, Canada*

Abstract

We list the datasets used for the empirical analysis in the CIVEMSA 2018 paper entitled *"Performance Analysis of Granular versus Traditional Neural Network Classifiers: Preliminary Results"*

We employed 123 well-known pattern classification datasets taken from the KEEL [1] and UCI Machine Learning (ML) [2] repositories. These datasets exhibit different characteristics and allow evaluating the predictive capability of both granular and non-granular neural classification models under consideration.

Table 1 outlines the number of instances, attributes and decision classes for each dataset. The presence of noise and the imbalance ratio (calculated as the ratio of the size of the majority class to that of the minority class) are also given. In this paper, we say that a dataset is imbalanced if the number of instances belonging to the majority decision class is at least five times the number of instances belonging to the minority class.

Table 1: Characterization of the ML datasets adopted for the simulations.

Dataset	Instances	Attributes	Classes	Noisy	Imbalance
acute-inflammation	120	6	2	no	no
acute-nephritis	120	6	2	no	no
anneal	898	38	6	no	85:1
appendicitis	106	7	2	no	no
arrhythmia	452	262	13	no	122:1
audiology	226	69	24	no	57:1
australian	690	14	2	no	no
autos	205	25	7	no	22:1
balance-noise	625	4	3	yes	5:1
balance-scale	625	4	3	no	5:1
ballons	16	4	2	no	no
banana	5300	2	2	no	no

Continued on next page

Table 1 – *Continued from previous page*

Dataset	**Instances**	**Attributes**	**Classes**	**Noisy**	**Imbalance**
bank	4521	16	2	no	7:1
breast	277	9	2	no	no
bc-wisconsin-diag	569	31	2	no	no
bc-wisconsin-prog	198	34	2	no	no
bridges-version1	107	12	6	no	no
bridges-version2	107	12	6	no	no
car	1728	6	4	no	17:1
cardiotocography-10	2126	35	10	no	10:1
cardiotocography-3	2126	35	3	no	9:1
chess	3196	36	2	no	no
cleveland	297	13	5	no	12:1
collins	500	23	15	no	13:1
contact-lenses	24	4	3	no	no
contraceptive	1473	9	3	no	no
credit-a	690	15	2	no	no
credit-g	1000	20	2	no	no
crx	653	15	2	no	no
csj	653	34	6	no	no
cylinder-bands	540	39	2	no	no
dermatology	358	34	6	no	5:1
echocardiogram	131	11	2	no	5:1
ecoli	336	7	8	no	71:1
ecoli0	220	7	2	no	no
ecoli-0vs1	220	7	2	no	no
ecoli1	336	7	2	no	no
ecoli2	336	7	2	no	5:1
ecoli3	336	7	2	no	8:1
ecoli-5an-nn	336	7	8	yes	71:1
eucalyptus	736	19	5	no	no
flags	194	28	8	no	15:1
glass	214	9	6	no	8:1
glass0	214	9	2	no	no
glass-0123vs456	214	9	2	no	no
glass1	214	9	2	no	no
glass-10an-nn	214	9	6	yes	8:1
glass2	214	9	2	no	no
glass-20an-nn	214	9	6	yes	8:1
glass3	214	9	2	no	6:1
glass-5an-nn	214	9	6	yes	8:1
glass6	214	9	2	no	6:1
hayes-roth	160	4	3	no	no
heart-5an-nn	270	13	2	yes	no
heart-statlog	270	13	2	no	no

Continued on next page

Table 1 – *Continued from previous page*

Dataset	**Instances**	**Attributes**	**Classes**	**Noisy**	**Imbalance**
horse-colic	368	22	2	no	no
horse-colic.orig	368	27	2	no	no
ionosphere	351	34	2	no	no
iris	150	4	3	no	no
iris0	150	4	2	no	no
iris-10an-nn	150	4	3	yes	no
iris-20an-nn	150	4	3	yes	no
iris-5an-nn	150	4	3	yes	no
labor	57	16	2	no	no
led7digit	500	7	10	no	no
libras	360	90	15	no	no
liver-disorders	345	6	2	no	no
lung-cancer	32	56	3	no	no
lymph	148	18	4	no	40:1
mammographic	830	5	2	no	no
mfeat-factors	2000	216	10	no	no
mfeat-fourier	2000	76	10	no	no
mfeat-karhunen	2000	64	10	no	no
mfeat-morpho	2000	6	10	no	no
mfeat-zernike	2000	47	10	no	no
molecular-biology	106	57	2	no	no
monk-2	432	6	2	no	no
mushroom	5644	22	2	no	no
musk-1	476	167	2	no	no
new-thyroid	215	5	2	no	5:1
optdigits	5620	64	10	no	no
ozone	2536	72	2	no	33:1
page-blocks	5473	10	5	no	175:1
parkinsons	195	22	2	no	no
phoneme	5404	5	2	no	no
pima	768	8	2	no	no
pima-10an-nn	768	8	2	yes	no
pima-20an-nn	768	8	2	yes	no
pima-5an-nn	768	8	2	yes	no
planning	182	12	2	no	no
postoperative	90	8	3	no	32:1
primary-tumor	339	17	22	no	841
saheart	462	9	2	no	no
segment	2310	19	7	no	no
solar-flare-1	323	5	6	no	11:1
sonar	208	60	2	no	no
soybean	683	35	19	no	11:1
spambase	4601	57	2	no	no

Continued on next page

Table 1 – *Continued from previous page*

Dataset	**Instances**	**Attributes**	**Classes**	**Noisy**	**Imbalance**
spectfheart	267	44	2	no	no
splice	3190	60	3	no	no
sponge	76	44	3	no	23:1
tae	151	5	3	no	no
tic-tac-toe	958	9	2	no	no
vehicle	846	18	4	no	no
vehicle0	846	18	2	no	no
vehicle1	846	18	2	no	no
vehicle2	846	18	2	no	no
vehicle3	846	18	2	no	no
vertebral2	310	6	2	no	no
vertebral3	310	6	3	no	no
vote	435	16	2	no	no
wall-following	5456	24	4	no	6:1
waveform	5000	40	3	no	no
weather	14	4	2	no	no
wine	178	13	3	no	no
wine-5an-nn	178	13	3	yes	no
winequality-red	1599	11	6	no	68:1
winequality-white	4898	11	7	no	439:1
wisconsin	683	9	2	no	no
yeast	1484	8	10	no	92:1
yeast1	1484	8	2	no	no
yeast3	1484	8	2	no	8:1
zoo	101	16	7	no	10:1

[1] J. Alcalá, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera, Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework, Journal of Multiple-Valued Logic and Soft Computing 17 (255-287) (2010) 11.

[2] M. Lichman, UCI machine learning repository (2013).
URL `http://archive.ics.uci.edu/ml`