Train In Data

Should we combine over- and under-sampling?

Comparison

A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data

Gustavo E. A. P. A. Batista
Ronaldo C. Prati
Maria Carolina Monard
Instituto de Ciências Matemáticas e de Computação
Caixa Postal 668, 13560-970
São Carlos - SP, Brazil
{gbatista, prati, mcmonard}@icmc.usp.br

Compared over-sampling and combined methods in 13 publicly available datasets

Table 4: AUC results for the original and over-sampled data sets.

Data set	Pruning	Original	Rand Over	Smote	Smote+Tomek	Smote+ENN
Pima	yes	81.53(5.11)	85.32(4.17)	**85.49(5.17)**	84.46(5.84)	83.66(4.77)
	no	82.33(5.70)	**86.03(4.14)**	85.97(5.82)	85.56(6.02)	83.64(5.35)
German	yes	79.19(5.84)	**84.65(3.80)**	80.74(5.43)	81.75(4.78)	80.91(4.36)
	no	**85.94(4.14)**	85.56(4.31)	84.51(4.55)	84.02(3.94)	83.90(3.70))
Post-operative	yes	49.29(2.26)	**68.79(23.93)**	55.66(24.66)	41.80(16.59)	59.83(33.91)
	no	**78.23(15.03)**	71.33(23.43)	68.19(26.62)	47.99(16.61)	59.48(34.91)
Haberman	yes	58.25(12.26)	71.81(13.42)	72.23(9.82)	75.73(6.55)	**76.38(5.51)**
	no	67.91(13.76)	73.58(14.22)	75.45(11.02)	**78.41(7.11)**	77.01(5.10)
Splice-ie	yes	98.76(0.56)	**98.89(0.47)**	98.46(0.87)	98.26(0.51)	97.97(0.74)
	no	**99.30(0.30)**	99.09(0.27)	99.19(0.28)	99.13(0.31)	98.88(0.34)
Splice-ei	yes	98.77(0.46)	98.80(0.44)	**98.92(0.44)**	98.87(0.44)	98.85(0.60)
	no	99.47(0.61)	**99.52(0.60)**	**99.52(0.26)**	99.51(0.32)	99.49(0.16)
Vehicle	yes	98.49(0.84)	**99.14(0.73)**	98.96(0.98)	98.96(0.98)	97.92(1.09)
	no	98.45(0.90)	**99.13(0.75)**	99.04(0.85)	99.04(0.85)	98.22(0.90)
Letter-vowel	yes	98.07(0.63)	98.80(0.32)	98.90(0.20)	98.90(0.20)	**98.94(0.22)**
	no	98.81(0.33)	98.84(0.27)	99.15(0.17)	99.14(0.17)	**99.19(0.15)**
New-thyroid	yes	94.73(9.24)	98.39(2.91)	98.91(1.84)	98.91(1.84)	**99.22(1.72)**
	no	94.98(9.38)	98.89(2.68)	98.91(1.84)	98.91(1.84)	**99.22(1.72)**
E.Coli	yes	87.64(15.75)	93.24(6.72)	95.49(4.30)	**95.98(4.21**	95.29(3.79)
	no	92.50(7.71)	93.55(6.89)	95.49(4.30)	**95.98(4.21**	95.29(3.79)
Satimage	yes	93.73(1.91)	95.34(1.25)	95.43(1.03)	95.43(1.03)	**95.67(1.18)**
	no	94.82(1.18)	95.52(1.12)	95.69(1.28)	95.69(1.28)	**96.06(1.20)**
Flag	yes	45.00(15.81)	**79.91(28.72)**	73.62(30.16)	79.30(28.68)	79.32(28.83)
	no	76.65(27.34)	79.78(28.98)	73.87(30.34)	**82.06(29.52)**	78.56(28.79)
Glass	yes	88.16(12.28)	92.20(12.11)	91.40(8.24)	91.40(8.24)	**92.90(7.30)**
	no	88.16(12.28)	92.07(12.09)	91.27(8.38)	91.27(8.38)	**93.40(7.61)**
Letter-a	yes	99.61(0.40)	99.77(0.30)	**99.91(0.12)**	**99.91(0.12)**	**99.91(0.12)**
	no	99.67(0.37)	99.78(0.29)	**99.92(0.12)**	**99.92(0.12)**	99.91(0.14)
Nursery	yes	99.79(0.11)	**99.99(0.01)**	99.21(0.55)	99.27(0.36)	97.80(1.07)
	no	99.96(0.05)	**99.99(0.01)**	99.75(0.34)	99.53(0.31)	99.20(0.51)

Scalability

- Methods based on KNN do not scale well
- Distance metrics
- Categorical variables

Train In Data

THANK YOU

www.trainindata.com