10th International Conference on Intelligent Data Engineering and Automated Learning
(IDEAL 2009)

Similarity-Binning Averaging: A Generalisation of Binning Calibration

Antonio Bella, Cèsar Ferri,
José Hernández-Orallo and
María José Ramírez-Quintana

Universitat Politècnica de València, Spain

Outline

- Introduction
- Traditional Calibration Methods
- Calibration by Multivariate Similarity-Binning Averaging
- Experimental Results
- Conclusions and Future Work

Introduction

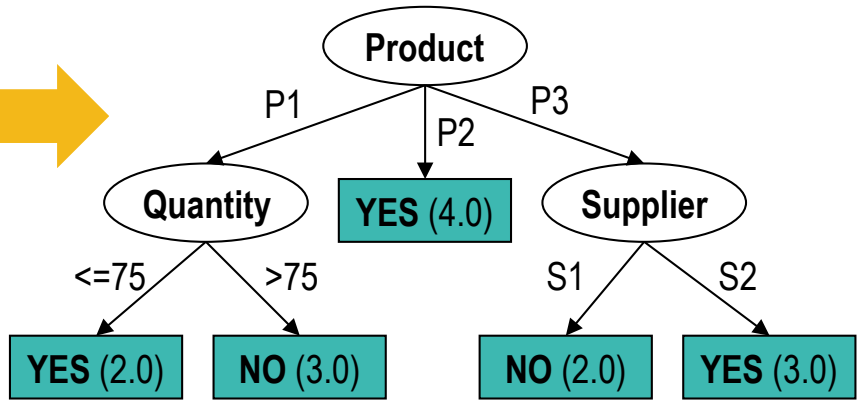
Introduction

Probabilistic estimation models

Training Data

Supplier	Product	Quantity	Price	Delivered on time?
S1	P1	85	85	NO
S2	P1	90	80	NO
S1	P2	86	83	YES
S1	P3	96	70	YES
S1	P3	80	68	YES
S2	P3	70	65	NO
S2	P2	65	64	YES
S1	P1	95	72	NO
S1	P1	70	69	YES
S1	P3	80	75	YES
S2	P1	70	75	YES
S2	P2	90	72	YES
S1	P2	75	81	YES
S2	P3	91	71	NO

Data Mining Model



New Data

Customer	Product	Quality	Price	Delivered on time?	Prob. (Yes)
S1	P1	70	70	YES	0.75
S2	P1	80	75	NO	0.2

Introduction

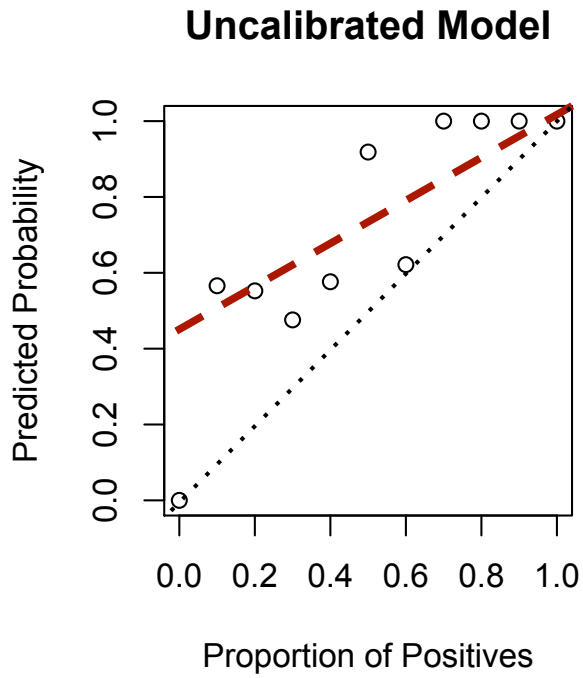
Why to calibrate?

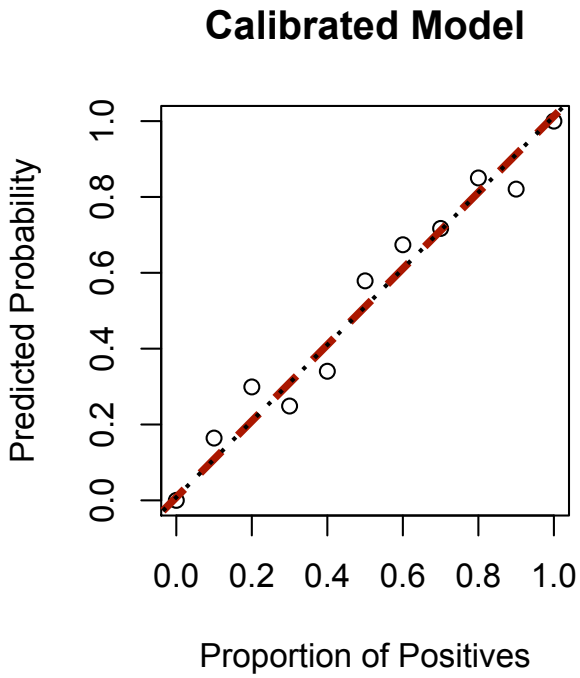
- Good in discriminating classes, but poorer in estimating probabilities.
- Instead of redesigning methods to obtain good probabilities, use calibration techniques.
- **Calibration technique:** post-processing method which aims at improving the probability estimation of a given classifier.

Introduction

Calibrated or uncalibrated?

- A classifier is calibrated if, for a sample of examples with predicted probability p, the expected proportion of positives is near to p.





Traditional Calibration Methods

Universitat Politècnica de València, Spain

Traditional Calibration Methods

- Binning averaging method.
- Pair-adjacent violators algorithm (PAV).
- Platt's method.

Traditional Calibration Methods

Common issues

- Based in ordering instances.
- Only binary problems (directly).
- Problem attributes are only used for calculating estimated probability.
- Estimated probability (of the positive class) is only used for ordering instances.
- All examples in a bin have the same calibrated probability.

Traditional Calibration Methods

Unexplored possibilities

- Probability calibration by similarity (*k*-most similar instances).
- Applicable to multiclass problems.
- Use estimated probabilities (of all the classes) and, also, the problem attributes for computing similarity between instances.
- More information can improve the calibrated probability.
- Each example has a calibrated probability.

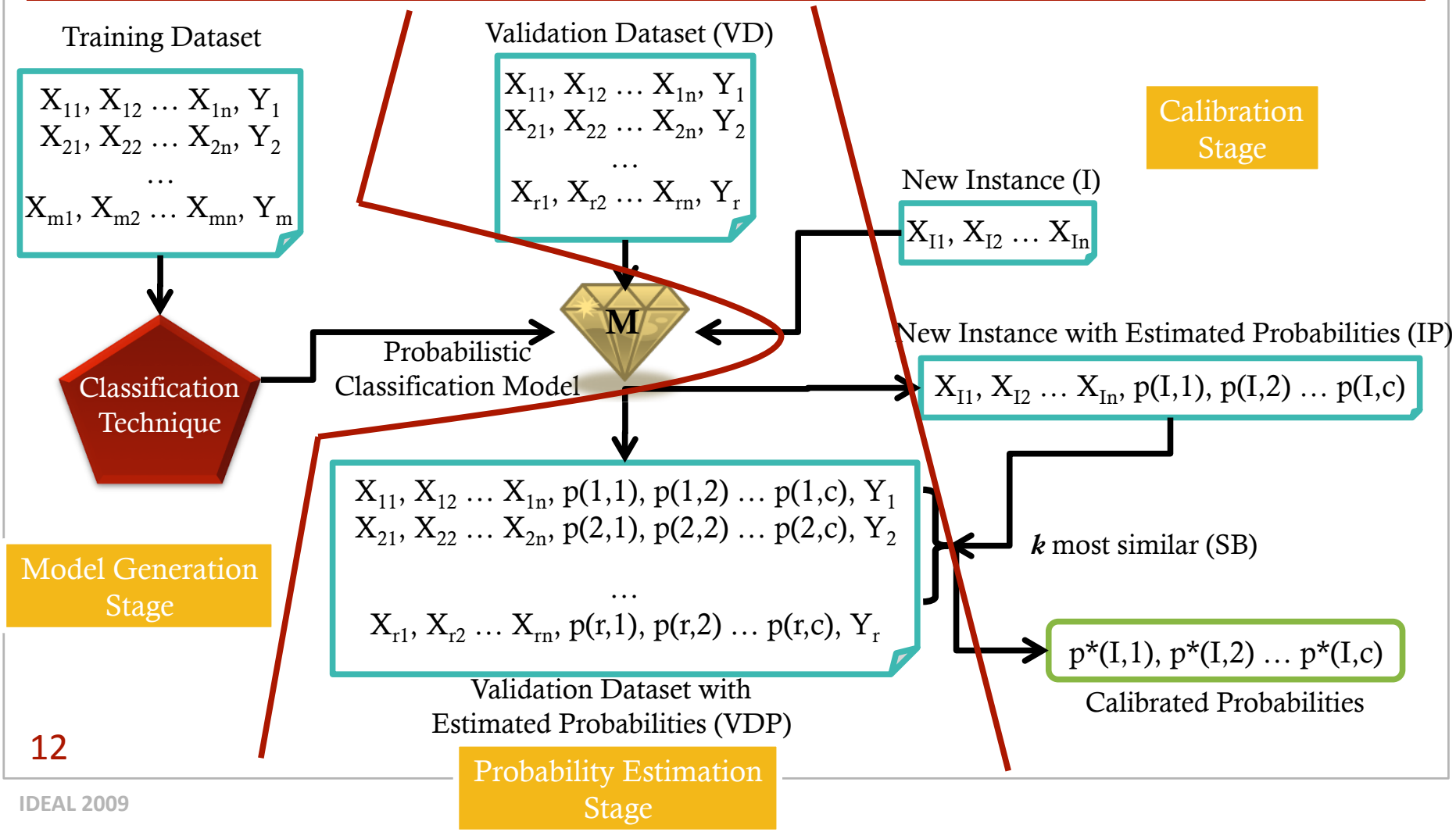
10th International Conference on Intelligent Data Engineering and Automated Learning
(IDEAL 2009)

Calibration by Multivariate Similarity-Binning Averaging

Universitat Politècnica de València, Spain

Calibration by Multivariate Similarity-Binning Averaging

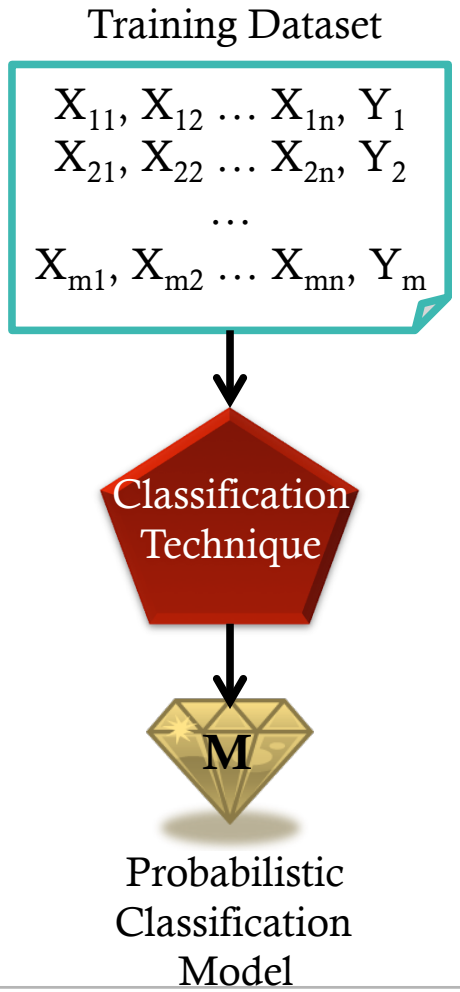
Schema



Calibration by Multivariate Similarity-Binning Averaging

Model generation stage

- Typical learning process.
- A classification technique is applied to a training dataset to learn a probabilistic classification model (*M*).
- This stage may not exist if the model is given beforehand (a hand-made model or an old model).



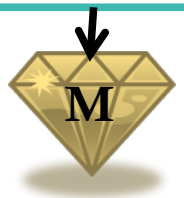
Calibration by Multivariate Similarity-Binning Averaging

Probability estimation stage

- The trained model *M* gives the estimated probabilities associated with a dataset.
- This dataset can be the same used for training, or an additional validation dataset *VD*.
- The estimated probability for each class is joined as new attribute, creating a new dataset *VDP*.

Validation Dataset (VD)

$X_{11}, X_{12} \ldots X_{1n}, Y_1$
$X_{21}, X_{22} \ldots X_{2n}, Y_2$
$\ldots$
$X_{r1}, X_{r2} \ldots X_{rn}, Y_r$



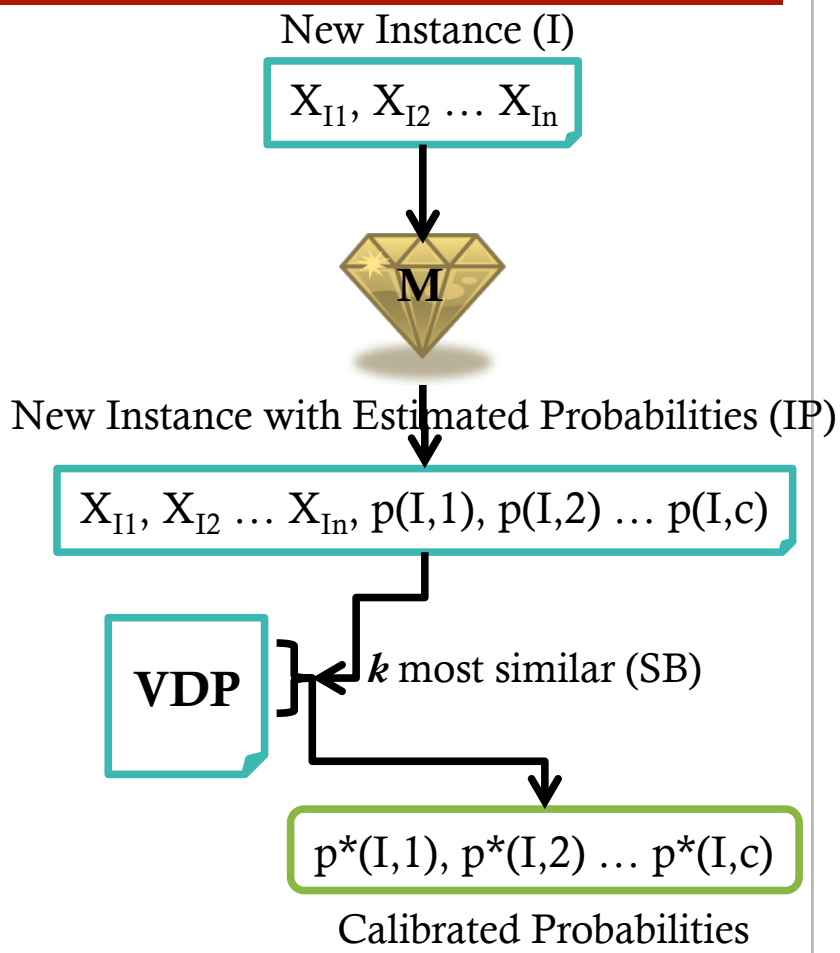
$X_{11}, X_{12} \ldots X_{1n}, p(1,1), p(1,2) \ldots p(1,c), Y_1$
$X_{21}, X_{22} \ldots X_{2n}, p(2,1), p(2,2) \ldots p(2,c), Y_2$
$\ldots$
$X_{r1}, X_{r2} \ldots X_{rn}, p(r,1), p(r,2) \ldots p(r,c), Y_r$

Validation Dataset with Estimated Probabilities (VDP)

Calibration by Multivariate Similarity-Binning Averaging

Calibration stage

- To calibrate a new instance *I*:
  1. Obtain estimated probabilities from the classification model *M*.
  2. Add these probabilities to the instance creating a new instance (*IP*).
  3. Select the *k*-most similar instances to this new instance from the dataset *VDP*.
  4. The calibrated probability of this instance *I* for each class is the predicted class probability of the *k*-most similar instances using all attributes.



Experimental Results

Experimental Results

Configuration

- 20 binary datasets from the UCI repository
- 2 different settings:
  - Training and test sets (75% / 25%)
  - Training, validation and test sets (56% / 19% / 25%)
- Classification techniques (WEKA):
  - Naïve Bayes, J48, IBk (k=10) and Logistic Regression
- Baseline methods:
  - Class: classification techniques without calibration
  - 10-NN: 10 most similar instances with the original attributes

Experimental Results

Calibration methods and measures

- Calibration methods:
  - Binning averaging (10 bins)
  - PAV algorithm
  - Platt's method
  - Similarity-Binning Averaging (SBA) (k=10)
- Calibration measures:
  - Calibration by overlapping bins (CalBin)
    - Pure calibration measure
  - Mean Squared Error (MSE)
    - Hybrid measure
    - Brier score decomposition
    - Calibration loss and refinement loss

Experimental Results

By dataset, CalBin measure

Dataset	ClassT	10-NNT	BinT	PAVT	PlattT	SBAT	BinV	PAVV	PlattV	SBAV
1	0.1953	0.1431	0.2280	0.2321	0.1856	0.1827	0.3092	0.2928	0.2446	0.1924
2	0.0494	0.0374	0.0647	0.0447	0.0623	0.0408	0.0791	0.0538	0.0775	0.0423
3	0.0698	0.1472	0.0501	0.0397	0.0434	0.0491	0.0548	0.0448	0.0479	0.0628
4	0.1517	0.1216	0.1533	0.1535	0.1421	0.1164	0.1996	0.1853	0.1563	0.1244
5	0.1220	0.0882	0.1060	0.1035	0.1132	0.0848	0.1408	0.1293	0.1269	0.0874
6	0.1250	0.1340	0.1263	0.1393	0.1268	0.1233	0.1933	0.1855	0.1227	0.1233
7	0.1192	0.1049	0.1220	0.1351	0.1205	0.1105	0.1889	0.1861	0.1267	0.1199
8	0.1984	0.2028	0.2316	0.2400	0.1998	0.1994	0.2877	0.2798	0.2777	0.2149
9	0.1476	0.1412	0.1690	0.1587	0.1529	0.1443	0.2247	0.1995	0.1834	0.1432
10	0.1632	0.1359	0.1643	0.1673	0.1727	0.1332	0.2082	0.1999	0.2597	0.1358
11	0.0665	0.0625	0.0777	0.0542	0.0791	0.0516	0.0945	0.0672	0.1006	0.0588
12	0.1380	0.1588	0.1179	0.0990	0.1358	0.1064	0.1701	0.1428	0.1854	0.1303
13	0.1876	0.2996	0.1984	0.1464	0.2110	0.1820	0.2914	0.1940	0.4478	0.2792
14	0.1442	0.2794	0.1618	0.1355	0.1046	0.1443	0.2067	0.1740	0.1340	0.1730
15	0.0395	0.0366	0.0418	0.0358	0.0468	0.0368	0.0434	0.0359	0.0494	0.0367
16	0.0296	0.0158	0.0270	0.0236	0.0250	0.0194	0.0297	0.0264	0.0285	0.0265
17	0.2606	0.1916	0.2343	0.2376	0.2374	0.2007	0.3207	0.2924	0.2750	0.1844
18	0.0945	0.0471	0.0636	0.0568	0.0951	0.0466	0.0733	0.0658	0.0964	0.0910
19	0.3138	0.3497	0.2995	0.2911	0.3110	0.3110	0.3615	0.3380	0.4265	0.3117
20	0.1240	0.2094	0.1260	0.1198	0.0906	0.0934	0.1736	0.1621	0.0971	0.0824
AVG.	0.1370	0.1453	0.1382	0.1307	0.1328	0.1188	0.1826	0.1628	0.1732	0.1310

Experimental Results

By dataset, MSE measure

Dataset	ClassT	10-NNT	BinT	PAVT	PlattT	SBAT	BinV	PAVV	PlattV	SBAV
1	0.2086	0.1912	0.2086	0.2095	0.2016	0.1998	0.2123	0.2136	0.2081	0.1982
2	0.0353	0.0262	0.0510	0.0343	0.0362	0.0306	0.0635	0.0375	0.0380	0.0316
3	0.0465	0.0648	0.0467	0.0387	0.0391	0.0227	0.0515	0.0424	0.0423	0.0332
4	0.1506	0.1351	0.1503	0.1449	0.1442	0.1336	0.1641	0.1535	0.1513	0.1371
5	0.1347	0.1121	0.1244	0.1203	0.1262	0.1160	0.1320	0.1239	0.1287	0.1176
6	0.1889	0.1795	0.1888	0.1883	0.1877	0.1814	0.1849	0.1832	0.1821	0.1800
7	0.1790	0.1749	0.1795	0.1786	0.1777	0.1821	0.1818	0.1779	0.1756	0.1835
8	0.1926	0.1992	0.1924	0.1936	0.1906	0.2005	0.1966	0.1982	0.1974	0.1957
9	0.1491	0.1435	0.1580	0.1503	0.1470	0.1469	0.1675	0.1534	0.1522	0.1390
10	0.1473	0.1294	0.1460	0.1483	0.1397	0.1305	0.1546	0.1505	0.1585	0.1271
11	0.0554	0.0568	0.0646	0.0482	0.0538	0.0456	0.0816	0.0574	0.0599	0.0524
12	0.1266	0.1297	0.1118	0.0981	0.1094	0.0996	0.1311	0.1071	0.1186	0.1141
13	0.1120	0.1233	0.1582	0.1128	0.1044	0.0907	0.2109	0.1419	0.2288	0.1196
14	0.1214	0.1047	0.1172	0.1030	0.1065	0.0517	0.1362	0.1164	0.1244	0.0819
15	0.0083	0.0006	0.0174	0.0040	0.0079	0.0001	0.0198	0.0045	0.0088	0.0003
16	0.0310	0.0311	0.0355	0.0266	0.0307	0.0241	0.0370	0.0275	0.0277	0.0370
17	0.2545	0.1760	0.2343	0.2286	0.2285	0.2080	0.2305	0.2157	0.2225	0.1847
18	0.1027	0.0814	0.0765	0.0721	0.0878	0.0690	0.0800	0.0746	0.0895	0.1042
19	0.2829	0.2459	0.2776	0.2637	0.2437	0.2692	0.2639	0.2482	0.2568	0.2276
20	0.1579	0.1141	0.1480	0.1448	0.1468	0.0817	0.1571	0.1522	0.1526	0.1108
AVG.	0.1343	0.1210	0.1343	0.1254	0.1255	0.1142	0.1428	0.1290	0.1362	0.1188

Experimental Results

Nemenyi statistical significance test (column vs. row)

10-NNT	BinT	PAVT	PlattT	SBAT	BinV	PAVV	PlattV	SBAV	CalBin
↑	→	=	→	↑	→	→	→	↑	ClassT
	→	→	→	=	→	→	→	=	10-NNT
		=	=	↑	→	→	→	↑	BinT
			→	↑	→	→	→	↑	PAVT
				↑	→	→	→	↑	PlattT
					→	→	→	=	SBAT
						=	↑	↑	BinV
							→	↑	PAVV
								↑	PlattV

(col. wins ↑, ties =, row wins →)

10-NNT	BinT	PAVT	PlattT	SBAT	BinV	PAVV	PlattV	SBAV	MSE
↑	→	=	=	↑	→	→	→	↑	ClassT
	→	↑	↑	↑	→	→	→	=	10-NNT
		↑	↑	↑	→	=	=	↑	BinT
			=	↑	→	→	→	=	PAVT
				↑	→	→	=	→	PlattT
					→	→	→	=	SBAT
						↑	↑	↑	BinV
							=	↑	PAVV
								↑	PlattV

Conclusions and Future Work

Conclusions

- New calibration method.
- Binning by constructing the bins using similarity to select the *k*-most similar instances (estimated probabilities and problem attributes).
- Experimental results show a significant increase in calibration for both measures considered, over three traditional calibration techniques.
- Can be applied to multiclass problems.

Future Work

- **Attribute-weighted *k*-NN** to form the bins in order to gauge the importance of attributes for cases when there is a great number of attributes or a great number of classes.
- **Locally-weighted *k*-NN**, where closer examples have more weight, in order to make the method more independent from *k*.
- Analysis of the method for **multiclass problems** comparing with other approximations, because binning, Platt's and PAV cannot deal directly with multiclass problems.

Thanks for your attention!

Antonio Bella
http://users.dsic.upv.es/~abella
abella@dsic.upv.es

Universitat Politècnica de València, Spain