Supplementary Material for MARLINE: Multi-Source Mapping Transfer Learning for Non-Stationary Environments

Honghui Du
School of Informatics
University of Leicester
Leicester, United Kingdom
hd168@leicester.ac.uk

Leandro L. Minku
School of Computer Science
University of Birmingham
Birmingham, United Kingdom
L.L.Minku@cs.bham.ac.uk

Huiyu Zhou
School of Informatics
University of Leicester
Leicester, United Kingdom
hz143@leicester.ac.uk

Appendix A
Complexity Analysis

MARLINE is composed of three main modules: Mapping Procedure, Centroids Update Procedure and Weighting Scheme. The time complexity for Mapping Procedure of one target example on one source concept is $\mathcal{O}(d^2)$. The time complexity for Update Centroids of each concept is $\mathcal{O}(d)$. The time complexity for the Weighting Scheme can be computed as follows:

- The time complexity of calculating projections of the target example is in:
$$\mathcal{O}((J_{S_1} + J_{S_2} + \cdots + J_{S_n} + J_T)d^2),$$
where $(J_{S_1} + J_{S_2} + \cdots + J_{S_n} + J_T)$ is the total number of base learner ensembles.
- The time complexity of Equations (11) and (12) is in:
$$\mathcal{O}((J_{S_1} + J_{S_2} + \cdots + J_{S_n} + J_T)K \times f_h),$$
where f_h is the time complexity of the sub-classifier to give a prediction and $(J_{S_1} + J_{S_2} + \cdots + J_{S_n} + J_T)K$ is the total number of sub-classifiers in the MARLINE ensemble.
- The time complexity of applying Equations (13) to (15) to all sub-classifiers is in:
$$\mathcal{O}((J_{S_1} + J_{S_2} + \cdots + J_{S_n} + J_T)K \times f_h)$$
- The time complexity of Equation (16) is in:
$$\mathcal{O}((J_{S_1} + J_{S_2} + \cdots + J_{S_n} + J_T)K)$$

The steps above are performed sequentially. Therefore, this gives us a Weighting Scheme time complexity in:

$$\mathcal{O}((J_{S_1} + J_{S_2} + \cdots + J_{S_n} + J_T)d^2 +$$
$$(J_{S_1} + J_{S_2} + \cdots + J_{S_n} + J_T)K \times f_h)$$

MARLINE's training procedure when the current training example is a target training example is composed of the drift detection method, the training of the base learner ensemble, centroids update and the weighting scheme, giving an overall time complexity of:

$$\mathcal{O}(f_{DD} + f_H +$$
$$(J_{S_1} + J_{S_2} + \cdots + J_{S_n} + J_T)d^2 +$$
$$(J_{S_1} + J_{S_2} + \cdots + J_{S_n} + J_T)K \times f_h),$$

where f_{DD} is the time complexity of the drift detection method and f_H is the training time complexity of the base learner ensemble algorithm.

When the training example is a source training example, only the drift detection method, the training of the base learner ensemble and the centroids update are needed, giving a lower overall time complexity of:

$$\mathcal{O}(f_{DD} + f_H + d)$$

MARLINE's prediction procedure is composed of the mapping procedure on all source concepts and all sub-classifiers' predictions, giving an overall time complexity of:

$$\mathcal{O}((J_{S_1} + J_{S_2} + \cdots + J_{S_n} + J_T)d^2 +$$
$$(J_{S_1} + J_{S_2} + \cdots + J_{S_n})K \times f_h$$

Note J_i used in the calculations above will be zero if no training example from data stream i has been produced yet.

APPENDIX B
PARAMETERS' VALUE

The grid search results of MARLINE are saved in a csv file provided alongside this report. The csv file follows the format shown in Table I.

TABLE I
GRID SEARCH RESULTS FORMAT

SourceType	Dataset	BLM	DDM	K	θ	σ

APPENDIX C
EXTRA EXPERIMENT RESULTS AND ANALYSES

Tables II, III, IV, V and VI show the mean accuracy and standard deviations of the experiment results across time steps. These were calculated based on 30 runs except for DWM, which is deterministic and requires a single run.

The Means and Standard Deviations of the accuracy across 30 runs are calculated as follows:

$$AverageAccuracy^t = \frac{\sum_{r=1}^{R} Accuracy_r^t}{R} \tag{1}$$

$$Mean = \frac{\sum_{t=1}^{t'} AverageAccuracy^t}{t'} \tag{2}$$

$$StandardDeviation = \sqrt{\frac{\sum_{t=1}^{t'}(AverageAccuracy^t - Mean)}{t'}} \tag{3}$$

where R is the number of runs with different random seeds and t' is the total number of time steps in the data stream.

From Tables II, III, IV and V, MARLINE achieved the largest increases in the mean accuracy of the base model when the class size was small (50) on the artificial datasets, which indicates that MARLINE has the ability to improve the predictive performance at the beginning of the data streams and right after the concept drifts. Moreover, the increase in accuracy of MARLINE over the base model was smaller for larger class sizes of the artificial datasets. This is because the learning problems of the artificial datasets are not so difficult to learn. Therefore, when the class size is medium (500) or large (5000), every method can reach good performance with time increasing. In this case, transfer learning will only help at the beginning of the data streams or after concept drifts. MARLINE achieved overall better performance due to its better predictive performance at the beginning of the data streams or after a concept drift.

For the real world datasets, MARLINE with or without sources always improved the mean accuracy over the base model. The probable reason for the results achieved by MARLINE is that the real world datasets have more complex learning problems (e.g the data stream is composed by difficult classification tasks) and concept drift situations, causing the base model to struggle to achieve good performance. In this case, transfer learning starts to help MARLINE to get better performance throughout time.

It is also worth noting that MARLINE obtained a better standard deviation than the other approaches in most cases, which means that MARLINE has a more smooth and stable performance, being more robust to concept drift.

TABLE II

MEAN ACCURACY AND STANDARD DEVIATION, WHERE BASE MODEL IS DDM+ONLINE BAGGING.

Dataset		CS/TSS	**MARLINE with source**	**MARLINE without source**	Melanie with source	Melanie without source	Base Model
No-Similar Source	No Drift	50	**0.913±0.1247**	0.8408±0.1406	0.7883±0.1406	0.7977±0.1841	0.7724±0.2007
		500	**0.9745±0.0364**	0.9651±0.0623	0.9623±0.0755	0.959±0.0877	0.9565±0.0896
		5000	**0.9908±0.0096**	0.9888±0.0211	0.9875±0.03	0.985±0.0335	0.9874±0.0303
	Abrupt	50	**0.9122±0.1032**	0.8858±0.1034	0.8367±0.1388	0.8275±0.1513	0.8291±0.1434
		500	**0.9779±0.0355**	0.94±0.0442	0.9667±0.0536	0.9639±0.0639	0.9612±0.066
		5000	**0.9898±0.0209**	0.9888±0.0232	0.9865±0.0293	0.9861±0.0319	0.9867±0.0265
	Incremental	50	**0.8404±0.1526**	0.8184±0.2124	0.7889±0.2192	0.7961±0.236	0.7983±0.2245
		500	0.882±0.1393	**0.8823±0.1377**	0.8729±0.1505	0.8667±0.1663	0.8682±0.1623
		5000	**0.8895±0.1291**	0.8876±0.1319	0.8697±0.1648	0.8655±0.1818	0.8757±0.1588
Similar Source	No Drift	50	0.9115±0.1246	0.8408±0.1406	**0.9433±0.1162**	0.7977±0.1841	0.7724±0.2007
		500	**0.9825±0.0234**	0.9651±0.0623	0.9812±0.0392	0.959±0.0877	0.9565±0.0896
		5000	**0.9902±0.0149**	0.9888±0.0211	0.9906±0.0161	0.985±0.0335	0.9874±0.0303
	Abrupt	50	**0.9394±0.097**	0.8858±0.1034	0.8893±0.1276	0.8275±0.1513	0.8291±0.1434
		500	**0.9791±0.0271**	0.94±0.0442	0.9748±0.0599	0.9639±0.0639	0.9612±0.066
		5000	**0.9908±0.0142**	0.9888±0.0232	0.9919±0.0147	0.9861±0.0319	0.9867±0.0265
	Incremental	50	0.8465±0.1647	0.8184±0.2124	**0.8753±0.1518**	0.7961±0.236	0.7983±0.2245
		500	0.8854±0.1358	0.8823±0.1377	**0.8962±0.1226**	0.8667±0.1663	0.8682±0.1623
		5000	0.8899±0.1301	0.8876±0.1319	**0.896±0.1234**	0.8655±0.1818	0.8757±0.1588
Real-World Data	Holiday	384	**0.8228±0.0622**	0.8085±0.0752	0.7337±0.0871	0.7558±0.1088	0.7236±0.1304
	Weekend	4970	**0.8489±0.0204**	0.8463±0.0238	0.7772±0.0286	0.7995±0.0373	0.7757±0.0473
	Weekday	12060	**0.8044±0.0553**	0.7414±0.0611	0.7237±**0.0491**	0.7234±0.0575	0.7059±0.0674

The values in red are the best values (not necessarily statistically speaking) in each row.

TABLE III

MEAN ACCURACY AND STANDARD DEVIATION, WHERE BASE MODEL IS DDM+ONLINE BOOSTING.

Dataset		CS/TSS	**MARLINE with source**	**MARLINE without source**	Melanie with source	Melanie without source	Base Model
No-Similar Source	No Drift	50	**0.913±0.1247**	0.8489±0.1407	0.8302±0.1332	0.807±0.1837	0.6693±0.2556
		500	**0.9698±0.0343**	0.9592±0.0593	0.9542±0.0808	0.949±0.0943	0.8918±0.1135
		5000	0.9859±**0.0109**	0.9873±0.0222	**0.9874**±0.027	**0.9874**±0.027	0.9733±0.0467
	Abrupt	50	**0.9291±0.096**	0.9041±0.1092	0.8712±0.1328	0.8844±0.1436	0.8361±0.1795
		500	**0.9839±0.029**	0.9787±0.0326	0.9749±0.0462	0.9725±0.0579	0.7897±0.1819
		5000	**0.9895±0.018**	0.9883±0.019	0.9888±0.0208	0.9887±0.0223	0.9824±0.0274
	Incremental	50	**0.8405±0.1581**	0.8087±0.1843	0.776±0.2147	0.8003±0.2028	0.7722±0.2322
		500	0.8781±0.1398	**0.8797±0.1394**	0.8266±0.2228	0.8499±0.1652	0.8533±0.1622
		5000	**0.8869±0.1307**	0.8868±0.1326	0.8646±0.1491	0.8743±0.1465	0.8785±0.1495
Similar Source	No Drift	50	0.913±0.1247	0.8489±0.1407	**0.9432±0.1162**	0.807±0.1837	0.6693±0.2556
		500	**0.9788±0.0247**	0.9592±0.0593	0.9775±0.0384	0.949±0.0943	0.8918±0.1135
		5000	**0.9893**±0.0183	0.9873±0.0222	0.9891±**0.0162**	0.9874±0.027	0.9733±0.0467
	Abrupt	50	0.9132±**0.1005**	0.9041±0.1092	**0.9267**±0.1064	0.8844±0.1436	0.8361±0.1795
		500	**0.9814±0.0312**	0.9787±0.0326	0.9771±0.0509	0.9725±0.0579	0.7897±0.1819
		5000	0.9896±0.0178	0.9883±0.019	**0.9906±0.0166**	0.9887±0.0223	0.9824±0.0274
	Incremental	50	0.84±0.1746	0.8087±0.1843	**0.8712±0.1595**	0.8003±0.2028	0.7722±0.2322
		500	0.8798±0.1424	0.8797±0.1394	**0.8907±0.1292**	0.8499±0.1652	0.8533±0.1622
		5000	0.887±0.1334	0.8868±0.1326	**0.893±0.127**	0.8743±0.1465	0.8785±0.1495
Real-World Data	Holiday	384	0.8±0.0645	**0.8158±0.0605**	0.732±0.0936	0.7523±0.104	0.6934±0.1978
	Weekend	4970	0.8078±0.0416	**0.826±0.0257**	0.7102±0.0471	0.7594±0.0401	0.8003±0.0699
	Weekday	12060	**0.748**±0.054	0.6666±0.0941	0.622±0.0156	0.6843±**0.0329**	0.7297±0.0621

The values in red are the best values (not necessarily statistically speaking) in each row.

TABLE IV

MEANS AND STANDARD DEVIATIONS, WHERE BASE MODEL IS $HDDM_A$+ONLINE BAGGING.

Dataset		CS/TSS	**MARLINE with source**	**MARLINE without source**	Melanie with source	Melanie without source	Base Model
No-Similar Source	No Drift	50	**0.913±0.1247**	0.8333±0.1369	0.8249±0.1547	0.7904±0.188	0.7723±0.2008
		500	**0.976±0.0365**	0.9673±0.0598	0.9649±0.0731	0.9579±0.0884	0.9565±0.0896
		5000	**0.9912±0.0094**	0.9888±0.0211	0.985±0.0335	0.9877±0.0297	0.9874±0.0303
	Abrupt	50	**0.913±0.1038**	0.8723±0.1194	0.8353±0.1416	0.8348±0.1507	0.8163±0.1525
		500	**0.9756±0.0375**	0.8858±0.0525	0.9386±0.0652	0.9354±0.0733	0.9385±0.0729
		5000	**0.9906±0.0203**	0.9886±0.023	0.9878±0.0243	0.9874±0.0274	0.9868±0.0285
	Incremental	50	**0.8367±0.1606**	0.8263±0.1929	0.8029±0.2152	0.8057±0.2168	0.808±0.2069
		500	**0.8918±0.1272**	0.8913±0.1279	0.8827±0.136	0.8829±0.1371	0.885±0.1323
		5000	**0.8961±0.1253**	0.8954±0.1247	0.8909±0.1301	0.8928±0.1286	0.8946±0.1255
Similar Source	No Drift	50	0.913±0.1247	0.8333±0.1369	**0.944±0.1163**	0.7904±0.188	0.7723±0.2008
		500	0.9798**±0.0292**	0.9673±0.0598	**0.9802**±0.0391	0.9579±0.0884	0.9565±0.0896
		5000	0.9902**±0.0149**	0.9888±0.0211	**0.9905**±0.0163	0.9877±0.0297	0.9874±0.0303
	Abrupt	50	**0.9429±0.098**	0.8723±0.1194	0.9138±0.115	0.8348±0.1507	0.8163±0.1525
		500	**0.9799±0.0318**	0.8858±0.0525	0.9705±0.0627	0.9354±0.0733	0.9385±0.0729
		5000	0.9912**±0.014**	0.9886±0.023	**0.9916**±0.0147	0.9874±0.0274	0.9868±0.0285
	Incremental	50	0.8476±0.1971	0.8263±0.1929	**0.87±0.1573**	0.8057±0.2168	0.808±0.2069
		500	0.8917±0.1289	0.8913±0.1279	**0.8974±0.1211**	0.8829±0.1371	0.885±0.1323
		5000	0.8961±0.1246	0.8954±0.1247	**0.8964±0.1231**	0.8928±0.1286	0.8946±0.1255
Real-World Data	Holiday	384	**0.8352±0.0529**	0.8123±0.0771	0.7102±0.0913	0.7501±0.1121	0.7177±0.1283
	Weekend	4970	**0.8791±0.0171**	0.8662±0.0253	0.7802±0.0343	0.8052±0.037	0.7812±0.0397
	Weekday	12060	**0.8071**±0.0551	0.7834±0.0705	0.7346**±0.0454**	0.7392±0.0492	0.7162±0.0654

The values in red are the best values (not necessarily statistically speaking) in each row.

TABLE V

MEAN ACCURACY AND STANDARD DEVIATION, WHERE BASE MODEL IS $HDDM_A$+ONLINE BOOSTING.

Dataset		CS/TSS	**MARLINE with source**	**MARLINE without source**	Melanie with source	Melanie without source	Base Model
No-Similar Source	No Drift	50	**0.913±0.1247**	0.8489±0.1407	0.8224±0.1479	0.8206±0.1787	0.6693±0.2556
		500	**0.9699±0.0343**	0.9592±0.0593	0.9542±0.0808	0.949±0.0943	0.8912±0.1135
		5000	**0.989±0.0109**	0.9873±0.0222	0.9874±0.027	0.9874±0.027	0.9734±0.0467
	Abrupt	50	**0.9298±0.0961**	0.9112±0.1093	0.8712±0.1328	0.8844±0.1436	0.8361±0.1795
		500	**0.9837±0.0293**	0.9752±0.0351	0.9738±0.0469	0.9702±0.0585	0.798±0.189
		5000	**0.9903±0.018**	0.9886±0.019	0.9883±0.0196	0.9883±0.0223	0.9877±0.0241
	Incremental	50	**0.8313±0.1971**	0.8051±0.2088	0.7808±0.2327	0.799±0.2153	0.7668±0.233
		500	**0.8817±0.1368**	0.8807±0.137	0.8479±0.1558	0.8624±0.1548	0.8648±0.1422
		5000	**0.891±0.1278**	0.8888±0.131	0.8662±0.1523	0.8769±0.1462	0.8898±0.1301
Similar Source	No Drift	50	0.9127±0.1246	0.8489±0.1407	**0.9432±0.1161**	0.8206±0.1787	0.6693±0.2556
		500	**0.9785±0.0247**	0.9592±0.0593	0.9775±0.0384	0.949±0.0943	0.8912±0.1135
		5000	**0.9892**±0.0182	0.9873±0.0222	0.989**±0.0162**	0.9874±0.027	0.9734±0.0467
	Abrupt	50	0.9242**±0.1035**	0.9112±0.1093	**0.9264**±0.1053	0.8844±0.1436	0.8361±0.1795
		500	**0.9827±0.0306**	0.9752±0.0351	0.9739±0.0569	0.9702±0.0585	0.798±0.189
		5000	**0.9907**±0.0163	0.9886±0.019	0.9906**±0.0162**	0.9883±0.0223	0.9877±0.0241
	Incremental	50	0.824±0.2058	0.8051±0.2088	**0.8606±0.1705**	0.799±0.2153	0.7668±0.233
		500	0.8868±0.1339	0.8807±0.137	**0.8934±0.125**	0.8624±0.1548	0.8648±0.1422
		5000	0.8909±0.1296	0.8888±0.131	**0.8936±0.1263**	0.8769±0.1462	0.8898±0.1301
Real-World Data	Holiday	384	0.7959±0.0754	**0.8137±0.062**	0.6798±0.1108	0.7553±0.104	0.6833±0.1954
	Weekend	4970	0.832±0.0311	**0.8459±0.0155**	0.6803±0.0604	0.7563±0.0388	0.7955±0.0736
	Weekday	12060	**0.7555**±0.0479	0.7272±0.0927	0.6242**±0.0176**	0.6706±0.0338	0.7291±0.0601

The values in red are the best values (not necessarily statistically speaking) in each row.

TABLE VI

MEAN ACCURACY AND STANDARD DEVIATION OF OTHER BASELINES.

Dataset		CS/TSS	Online Bagging	Online Boosting	Adaptive Random Forest(DDM)	Adaptive Random Forest(HDDMA)	Dynamic Weighted Majority
No-Similar Source	No Drift	50	0.7724±0.2007	0.6693±0.2556	0.7859±0.1944	0.7859±0.1944	0.7926±0.1955
		500	0.9565±0.0896	0.8912±0.1135	0.9373±0.0907	0.9397±0.0886	0.971±0.0459
		5000	0.9874±0.0303	0.9734±0.0467	0.985±0.028	0.9848±0.0283	0.9901±0.0173
	Abrupt	50	0.8163±0.1525	0.8361±0.1795	0.8332±0.1485	0.8214±0.1539	0.8212±0.1508
		500	0.9365±0.0735	0.798±0.189	0.9528±0.0649	0.949±0.0757	0.9436±0.0549
		5000	0.9706±0.0403	0.9877±0.0241	0.9851±0.0298	0.9803±0.0335	0.9779±0.037
	Incremental	50	0.5802±0.3275	0.7498±0.2155	0.8016±0.2141	0.7782±0.2415	0.8016±0.207
		500	0.5555±0.362	0.8254±0.1705	0.852±0.1849	0.8578±0.1725	0.8605±0.1818
		5000	0.7666±0.2651	0.8423±0.1739	0.8713±0.1589	0.8823±0.1415	0.8932±0.1265
Similar Source	No Drift	50	0.7724±0.2007	0.6693±0.2556	0.7859±0.1944	0.7859±0.1944	0.7926±0.1955
		500	0.9565±0.0896	0.8912±0.1135	0.9373±0.0907	0.9397±0.0886	0.971±0.0459
		5000	0.9874±0.0303	0.9734±0.0467	0.985±0.028	0.9848±0.0283	0.9901±0.0173
	Abrupt	50	0.8163±0.1525	0.8361±0.1795	0.8332±0.1485	0.8214±0.1539	0.8212±0.1508
		500	0.9365±0.0735	0.798±0.189	0.9528±0.0649	0.949±0.0757	0.9436±0.0549
		5000	0.9706±0.0403	0.9877±0.0241	0.9851±0.0298	0.9803±0.0335	0.9779±0.037
	Incremental	50	0.5802±0.3275	0.7498±0.2155	0.8016±0.2141	0.7782±0.2415	0.8016±0.207
		500	0.5555±0.362	0.8254±0.1705	0.852±0.1849	0.8578±0.1725	0.8605±0.1818
		5000	0.7666±0.2651	0.8423±0.1739	0.8713±0.1589	0.8823±0.1415	0.8932±0.1265
Real-World Data	Holiday	384	0.7076±0.1377	0.6936±0.1972	0.7634±0.1057	0.7481±0.1149	0.7885±0.0987
	Weekend	4970	0.7777±0.0487	0.7971±0.0713	0.8457±0.0283	0.839±0.0306	0.7882±0.0409
	Weekday	12060	0.7118±0.062	0.7257±0.0686	0.7569±0.0482	0.7612±0.0517	0.7201±0.0649