Discrete Source Classifier: Benchmarking Experiments with Cycle 2A Data

prepared by:	C. Elting, C. A. L. Bailer-Jones
approved by:	
reference:	GAIA-C8-TN-MPIA-CE-001-01
issue:	1
revision:	1
date:	15 OCT 2007
status:	Issued

Abstract

This documents describes DSC benchmarking experiments using cycle 2A data. We investigated classification algorithms based on boosting, neural networks, mixture clustering and radial basis function networks. Two different magnitudes were used (17 and 20). Moreover classification was done with astrometry as well as without.

Document History

Issue	Revision	Date	Author	Comment
D	0	15-oct-2007	CE, CBJ	First draft

Contents

1 Acronyms

The following is a list of the acronyms used in this document.

Acronym	Description
ANN	Artificial Neural Network
BIC	Baysian Information Criterion
CART	Classification and Regression Trees
CPU	Central Processor Unit
DSC	Discrete Source Classifier
EM	Engineering Model
GB	GigaByte
PCA	Principal Component Analysis
RAM	Random Access Memory
RBF	Radial Basis Function
RT	Real-Time
SVM	Support Vector Machine

2 Introduction

2.1 Data and hardware

For the tests we used noisy cycle 2A data for magnitudes $G = 17$ and $G = 20$ which are described in detail in sections 3.1 and 4.1. We also conducted experiments in which stars and physical binaries were merged into a single category (”stellar”). The data were mean-scaled prior to the experiments.

All tests were conducted on a SUSE Linux 10.1 server with an Intel Xeon CPU with 2.80 GHz and 3.71 GB RAM. All tests were conducted using the R software package version 2.2.0 (R Project (2007)).

2.2 Boosting

For boosting we used the Adaboost.M1 algorithm from Cortés et al. (2007) which is implemented by the `adabag` library in R. The basic idea of boosting is to combine several weak classifiers instead of using one single classifier. The Adaboost.M1 algorithm uses classification and regression trees (CART) as the basic classifier. The CART results are combined by means of weights which are learned iteratively by means of the expectation maximisation algorithm (EM algorithm).

In the `adabag` implementation we varied the number of iterations of the EM algorithm. For the tests a maximum CART depth equal to the number of classes in the experiment (either three or four) was used unless otherwise noted.

2.3 Mixture Clustering with PCA

We also investigated mixture clustering which is implemented by the R library `mclust` as detailed in Fraley & Raftery (2006). For every class given in the training data mixture clustering conducts a hierarchical clustering. This results in K superclusters where K is the number of classes used. Afterwards to each subcluster of every supercluster a Gaussian mixture component is fit. Each of the K classes is then modeled as the weighted sum of its mixture components.

Moreover by using the Baysian information criterion (BIC) the best covariance model for all mixture components within one supercluster is computed. This model is denoted by one to three characters (e.g., “VVV”), which represent a model for the class-dependent covariance matrices of the mixture components (see Fraley & Raftery (2006) for more information). In the algorithm the number of mixture components for each supercluster may be set to an interval instead of a fixed number. In this case the algorithm determines the best suitable number of mixture components for each supercluster.

For higher numbers of mixtures numeric difficulties arise when using the full set of 98 pixels (resp. 96 pixels when using BP/RP only). Therefore a principal component analysis (PCA) was conducted using the `prcomp` function in R prior to the mixture clustering. The number of mixtures to be used was varied between interval $[1, 1]$ and interval $[1, 10]$. The number of principal components was varied during the experiment between 1 (maximum data reduction) and 98 resp. 96 (no data reduction).

2.4 Neural Networks

Neural networks are a non-linear classification method, which learns relations between input nodes (e.g., BP/RP data and astrometry) and output nodes (e.g., the source type). In the multilayer preceptron architecture one or more hidden layers are introduced which are situated between the input layer and the output layer. All nodes between neighbouring layers are connected. The number of nodes in the hidden layer is arbitrary and depends on the nature of the classification problem. For each edge weights are learned by means of an iterative algorithm. These weights are used in connection with sigmoidal transfer functions when the neural network is applied to data. Moreover weight decay is used to prevent overfitting of the data during training.

We investigated neural networks using a single hidden layer. To do so we were using the `nnet` library of R Venables & Ripley (2002). We varied the size of the hidden layer between 1 and 9 nodes. Moreover we used a weight decay parameter, for which values between 0 and 100 were used. We used a maximum of 1000 training iterations for all experiments.

2.5 Radial Basis Function Networks

Radial basis function networks are similar to neural networks as they also consist of an input layer, one hidden layer and an output layer. However instead of the sigmoidal activation function used in neural networks radial basis functions are used, which typically are Gaussians. Mean and standard deviation of the Gaussians are learned by means of an unsupervised k-means clustering algorithm.

For tests with radial basis function networks we used the `RWeka` library. This library uses an R interface to the Java-based Weka machine learning library. Weka implements a radial basis function network using normalised Gaussians. It uses the k-means clustering algorithm to provide the basis functions and learns a logistic regression on top of that.

We varied the number of basis functions which equals the number of nodes in the hidden layer of the radial basis function network. Moreover we varied the regularisation parameter λ for the logistic regression. More information on the class `RBFNetwork` which implements radial basis function networks can be found in the corresponding Java documentation for this class[1].

3 Magnitude 17

3.1 Data sets

The features of the data set comprise 96 concatenated BP/RP bins as well as two features for parallaxes as well as proper motions. The data set was randomly divided into a training set and an evaluation set (table 2). The data used are available under Subversion in the MPIA development directory (see Data file for $G = 17$ training; Data file for $G = 17$ evaluation in bibliography). The physical binaries used in the data had a brightness ratio of between zero and four. Moreover we added end-of-mission noise which was based on 80 transits.

[1] http://weka.sourceforge.net/doc/weka/classifiers/functions/RBFNetwork.html

Table 2: Data sets for $G = 17$ data.

Training set			
Galaxies	Physical Binaries	Quasars	Stars
1 987	1 998	2 237	2 147

Evaluation set			
Galaxies	Physical Binaries	Quasars	Stars
1 988	1 998	2 238	2 146

3.2 BP/RP only

3.2.1 Adaboost.M1-tests with adabag

For boosting the Adaboost.M1-implementation of the `adabag` R-library was used. Table 3 shows the total classification error for different iterations of the EM algorithm. Additionally the runtimes (in real time) for the training (RT Train) and the evaluation (RT Eval) were measured.

Table 3: Adaboost.M1 results for $G = 17$ data using BP/RP only: Total classification error and runtimes in real time.

Iterations	Total Error	RT Train	RT Eval
	%	sec	sec
1	42.41	18.6	1.2
100	37.81	1 592.8	79.9
200	37.82	3 267.5	200.1
300	37.91	4 902.2	362.5
400	37.83	6 550.3	571.8
500	37.85	8 321.5	857.0
600	37.87	10 099.5	1 183.4

The lowest total classification error (37.81%) was obtained for 100 iterations. For this result table 4 shows the confusion matrix. The first column contains the true values of the observations. Columns two to five contain the estimated classes. Each row sums up to 100%. The confusion matrix shows that most misclassifications occurred for stars of which less than a third were classified correctly. This is also true for the confusion matrices in experiments with more than 100 iterations which produced similar matrices.

Table 4: Adaboost.M1 results for $G = 17$ data using BP/RP only: 100 iterations: Confusion matrix in percent.

True classes	Galaxy	Phys. Binary	Quasar	Star
Galaxy	65.49	29.23	5.28	0.00
Phys. Binary	19.42	74.47	1.05	5.06
Quasar	8.36	2.46	79.89	9.29
Star	27.91	39.56	3.31	29.22

3.2.2 Neural Networks

For neural networks we used the `nnet` library of R R Project (2007). `nnet` implements a multi-layer perceptron with one hidden layer. We varied the size of the layer as well as the value of decay for the weights of the neural network (table 5). We chose values between 0 and 100 for the weight decay (cf., Venables & Ripley (2002)). The maximum number of iterations was 1 000.

Table 5: Neural network results for $G = 17$ data using BP/RP only: Total classification error [%].

Size	Decay								
	0	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	100
1	42.04	41.85	54.40	55.63	41.78	42.13	47.04	55.77	81.29
2	38.12	29.46	32.32	29.73	29.47	29.53	31.70	37.51	56.38
3	24.72	24.48	21.57	20.38	20.69	19.63	21.76	29.21	54.97
4	25.65	20.50	18.65	18.18	19.31	16.80	18.72	28.60	53.06
5	18.08	23.38	21.52	19.55	14.04	14.10	14.41	25.99	51.66
6	23.54	15.15	24.40	14.71	12.04	11.59	14.18	26.75	51.00
7	19.61	18.78	17.99	13.00	14.84	10.97	12.89	24.10	51.48
8	17.69	17.35	14.97	16.09	12.35	9.99	12.66	25.01	50.57
9	17.93	17.60	18.65	15.59	13.06	9.44	11.88	24.05	50.92

The best result (9.44%) was obtained using $size = 9$ and $decay = 10^{-1}$. For this experiment the confusion matrix is shown in table 6. The runtime for the training was 508.2 sec (real time). The runtime for the predictions was 1.1 sec.

Table 6: Neural network results for $G = 17$ data using BP/RP only, $size = 9$, $decay = 10^{-1}$: Confusion matrix in percent.

True classes	Galaxy	Phys. Binary	Quasar	Star
Galaxy	97.38	1.91	0.10	0.60
Phys. Binary	0.90	84.99	0.05	14.06
Quasar	0.89	0.09	96.74	2.28
Star	1.03	14.49	1.49	82.99

The results show that using neural networks resulted in a considerably lower classification error that when using the Adaboost.M1 algorithm (37.81%, section 3.2.1). Similar to Adaboost.M1 most misclassifications occurred for stars. As in the Adaboost.M1 experiments a considerable amount of stars was classified as a physical binary. Moreover roughly the same amount (14%) of physical binaries were classified as stars. Therefore the misclassifications between stars and physical binaries account for most of the total classification error of neural networks.

3.2.3 Mixture clustering with PCA

The maximum number N of mixture components per class was varied between 1 and 10, i.e., the mixture components were integer elements of the interval $[1, N]$, $1 \leq N \leq 10$. The number of principal components was also varied during the experiment between 1 (maximum data reduction) and 96 (no data reduction). The results are shown in table 7. Compared to the three class case (section 3.3.3) the dimension reduction of the PCA seems to have a bigger impact. The usage of 40 components instead of 80 resulted in an error reduction from 16.48% to 11.17% when using up to 9 mixture components per class.

For higher numbers of principal components numeric instabilities like singular covariance matrices took place which were due to the number of parameters to be learned. Experiments where those instabilities occurred are denoted by a dash in table 7.

Table 7: Mixture clustering with PCA results for $G = 17$ data using BP/RP only: Total classification error [%].

Mixtures	Number of PCs										
	1	10	20	30	40	50	60	70	80	90	96
$[1, 1]$	59.24	35.21	27.47	22.82	21.90	21.81	21.73	21.81	21.95	21.96	21.84
$[1, 2]$	57.87	26.32	19.86	16.67	16.33	16.01	16.37	16.65	16.33	17.60	19.45
$[1, 3]$	57.50	22.80	15.73	13.27	12.66	14.79	15.87	14.97	15.10	17.05	18.98
$[1, 4]$	57.50	22.40	15.37	13.33	11.77	14.13	15.63	14.79	15.20	18.58	19.16
$[1, 5]$	56.33	18.17	12.81	11.62	11.35	13.45	14.08	14.76	14.33	16.29	18.98
$[1, 6]$	56.33	17.41	12.83	10.93	11.30	13.45	14.09	14.77	14.34	16.57	–
$[1, 7]$	56.33	15.97	12.81	11.24	11.30	13.45	14.09	–	16.14	16.70	–
$[1, 8]$	56.33	15.84	12.81	11.24	11.30	13.45	14.08	–	16.23	16.47	–
$[1, 9]$	56.33	15.04	12.62	11.17	11.30	13.45	14.08	–	16.24	16.48	–
$[1, 10]$	56.33	15.05	–	–	11.30	13.45	–	–	–	–	–

The best result was a total error rate of 10.93% which was obtained using 30 principal components and up to 6 mixture components per class. Using more principal components did increase the error which might be due to the fact that a better fitting of the data was possible when using fewer dimensions. Moreover using more mixture components per class did neither decrease the error rate. The confusion matrix for this result is shown in table 8. The runtime for the training was 736.6 sec (real time). The runtime for the evaluation on the test set was 2.4 sec.

Table 8: Mixture clustering with PCA results for $G = 17$ data using BP/RP only, up to 6 mixture components per class, 30 principal components: Confusion matrix in percent.

True classes	Galaxies	Phys. Binary	Quasars	Stars
Galaxy	98.89	0.96	0.05	0.10
Phys. Binary	0.15	86.99	0.05	12.81
Quasars	0.00	0.13	97.90	1.97
Stars	0.23	25.96	1.13	72.69

Similar to Adaboost.M1 (section 3.2.1) and neural networks (section 3.2.2) most misclassifications were attributed to confusions between stars and physical binaries. The total classification error was slightly higher than for neural networks. For the same experiment table 9 shows the covariance models and the number of mixtures per class which were determined during training. The table shows that the maximum of 6 mixture components were used to model stars whereas fewer mixture components were used to model the other classes. Moreover the most flexible covariance model “VVV” was chosen for the mixture components in every class (arbitrary volume, shape and orientation of the covariance ellipsoid, see Fraley & Raftery (2006)).

Table 9: Mixture clustering with PCA results for $G = 17$ data using BP/RP only, up to 6 mixture components per class, 30 principal components: Covariance models and number of mixture components per class.

	Galaxies	Phys. Binary	Quasars	Stars
Covariance Model	VVV	VVV	VVV	VVV
#Mixtures	4	4	5	6

For stars there seemed to be a tendency of using more mixture components that for the other classes. When using 30 principal components and a maximum of 9 mixtures all of the 9 mixtures components were used to model stars. However only a maximum of six mixture components was used to model any of the other classes. This might be an indication that stars needed a more complex modelling than the other classes.

3.2.4 Radial basis function networks

For radial basis function networks we used the `RWeka` library. We varied the number of basis functions between 1 and 250 as well as the regularisation parameter for the logistic regression between 0 and 5 (table 10). The results show that after 100 basis functions a saturation effect takes place for the total classification error which is not decreasing further for higher numbers of radial basis functions. The results in the table also suggest that the decay has only little effect on the total classification error and does not lead to any significant improvements.

Table 10: RBF network results for $G = 17$ data using BP/RP only: Total classification error [%].

Size	Decay					
	0	1	2	3	4	5
1	50.26	50.36	50.35	50.35	50.37	50.37
50	39.58	40.06	40.11	40.13	40.13	40.14
100	38.52	38.72	38.79	38.78	38.83	38.84
150	39.07	39.12	39.12	39.12	39.12	39.16
200	39.64	39.76	39.75	39.74	39.75	39.71
250	39.75	39.27	39.27	39.22	39.09	39.04

The confusion matrix for the best result (38.52%) using 100 basis functions and no decay is shown in table 11. Compared to neural networks and mixture clustering (sections 3.2.1 and 3.2.2) the total classification error is significantly higher. Most misclassifications seemed to take place for stars of which two thirds were misclassified. The runtime for the training was 1 257.9 sec (real time). The runtime for the evaluation was 142.3 sec.

Table 11: RBF network results for $G = 17$ data using BP/RP only, $size = 100$, $decay = 0$: Confusion matrix in percent.

True classes	Galaxy	Phys. Binary	Quasar	Star
Galaxy	71.03	20.93	5.03	3.02
Phys. Binary	23.37	65.72	1.55	9.36
Quasar	10.59	2.37	76.14	10.90
Star	28.24	33.41	4.94	33.41

3.3 BP/RP only: Stars and physical binaries merged

As many misclassifications in section 3.2 were attributed to misclassifications between stars and physical binaries we also conducted experiments in which the data for both classes were merged into a single category named “stellar”. We expected a signification reduction of the total classication error as a result of this merger.

3.3.1 Adaboost.M1-tests with adabag

Merging stars and physical binaries resulted in an error reduction of about 3% (cf., section 3.2.1) for the Adaboost.M1 algorithm. The results for different iterations of the EM algorithm are displayed in table 12.

Table 12: Adaboost.M1 results for $G = 17$ data using BP/RP only with stars and physical binaries merged.

Iterations	Total Error %	RT Train sec	RT Eval sec
1	82.59	12.62	1.17
100	34.29	1 155.57	80.70
200	33.92	2 349.28	191.15
300	34.34	3 620.28	373.28
400	34.06	4 983.56	569.75
500	34.15	6 050.63	819.75
600	33.94	7 449.70	1 161.20

The lowest total classification error (33.92%) was obtained for 200 iterations. Higher numbers of iterations did not lead to a further reduction of the total classification error. For 200 iterations table 13 shows the confusion matrix. It is surprising that nearly all of the galaxies were misclassified as stellar (cf. experiments using all four classes, section 3.2.1). On the other hand nearly every stellar object is correctly classified as expected. Compared to using four classes more quasars were misclassified as stellar too.

It should also be noted that due to the merger of both categories there are roughly twice as many stellar objects as galaxies. Therefore recognising nearly all of the stellar objects has a greater impact on the total classification error than failing to recognise nearly every galaxy object.

Table 13: Adaboost.M1 results for $G = 17$ data using BP/RP only with stars and physical binaries merged: 200 iterations: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	0.45	0.15	99.40
Quasar	2.01	62.96	35.03
Stellar	0.19	0.56	99.25

3.3.2 Neural Networks

Again we varied the size of the hidden layer between 1 and 9 and the weight decay between 0 and 100. Table 14 shows the total classification error.

Table 14: Neural network results for $G = 17$ data using BP/RP only with stars and physical binaries merged: Total classification error [%].

Size	Decay								
	0	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	100
1	38.22	93.92	28.77	84.91	84.89	26.02	92.11	90.62	84.72
2	23.67	17.93	16.79	16.85	17.62	15.68	13.09	22.22	41.74
3	21.66	18.91	12.10	7.30	7.98	7.91	9.03	16.13	42.26
4	10.02	9.53	9.30	6.38	3.88	5.05	7.01	13.41	41.70
5	9.24	7.30	6.91	6.87	5.29	3.33	5.08	12.75	42.14
6	9.33	6.71	8.41	4.78	4.29	2.43	4.27	11.28	41.39
7	8.79	5.48	7.65	5.91	4.00	2.41	3.99	12.03	41.21
8	6.94	8.21	8.11	2.87	3.29	1.83	3.78	10.91	40.93
9	7.18	6.59	4.55	2.93	1.92	2.02	3.44	11.09	40.81

The best result (1.83%) was obtained using $size = 8$ and $decay = 10^{-1}$. This corresponds to the best result in the experiments using four classes which used $size = 9$ and $decay = 10^{-1}$ (section 3.2.2). Compared to this result the error rate has been reduced by 7.61% by merging stars and physical binaries.

For the best result the confusion matrix is shown in table 15. The runtime for the training was 345.1 sec (real time). The runtime for the predictions was 1.1 sec. As a difference to experiments with Adaboost.M1 (section 3.3.1) the merger did not increase the number of misclassifications of galaxies or quasars.

Nearly every object which was classified as being a physical binary or a star when using four classes was classified as being a stellar object when using three classes only (cf. section 3.2.2). Therefore using neural networks on $G = 17$ data with stars and physical binaries merged resulted in a near perfect classification.

Table 15: Neural network results for $G = 17$ data using BP/RP only, $size = 9$, $decay = 10^{-1}$: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	98.49	0.10	1.41
Quasar	1.43	96.74	1.83
Stellar	0.53	0.68	98.79

3.3.3 Mixture clustering with PCA

The maximum number of mixture components per class was varied between 1 and 10. The number of principal components was also varied during the experiment between 1 (maximum data reduction) and 96 (no data reduction).

Table 16: Mixture clustering with PCA results for $G = 17$ data using BP/RP only with stars and physical binaries merged: Total classification error [%].

Mixtures	Number of PCs										
	1	10	20	30	40	50	60	70	80	90	96
$[1, 1]$	57.69	32.75	22.28	12.61	8.51	8.09	7.90	7.84	8.20	8.18	8.24
$[1, 2]$	48.76	14.10	8.60	6.20	5.61	5.35	5.42	5.54	5.26	6.07	6.01
$[1, 3]$	48.14	9.64	4.07	3.15	2.80	2.60	3.57	2.21	2.58	2.58	5.09
$[1, 4]$	48.14	7.81	3.11	2.60	1.77	1.98	3.26	2.03	2.69	4.76	5.07
$[1, 5]$	48.39	6.14	2.28	1.72	0.81	1.32	1.84	1.95	1.92	2.17	5.01
$[1, 6]$	48.39	5.35	1.96	1.36	0.80	1.33	1.83	1.96	1.92	2.17	5.03
$[1, 7]$	48.39	4.72	1.76	1.48	0.80	1.32	1.84	1.96	1.92	2.17	5.03
$[1, 8]$	48.38	4.22	1.76	1.09	0.80	1.33	1.84	–	1.92	2.17	4.48
$[1, 9]$	48.38	3.60	1.64	1.08	0.80	1.32	1.84	–	1.92	2.17	5.41
$[1, 10]$	48.39	3.30	–	1.08	0.80	1.33	–	–	–	–	–

The best result was a total error rate of 0.80% which was obtained using 40 principal components and up to 6 mixture components per class. Using more principal components did increase the error which might be due to the fact that a better fitting of the data was possible when using fewer dimensions. Moreover using more mixture components per class did neither decrease the error rate. The confusion matrix for this result is shown in table 17. The runtime for the training was 1 302.6 sec (real time). The runtime for the evaluation on the test set was 2.2 sec.

Table 17: Mixture clustering with PCA results for $G = 17$ data using BP/RP only with stars and physical binaries merged, up to 6 mixture components per class, 40 principal components: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	99.60	0.00	0.40
Quasar	0.00	98.53	1.48
Stellar	0.24	0.39	99.37

Similar to neural networks (section 3.3.2) merging stars and physical binaries had no negative effect on the total classification error regarding galaxies and quasars and resulted in a near perfect classification. For the same experiment table 18 shows the covariance models and the number of mixtures per class which were determined during training. The number of mixture components used to model galaxies or quasars did not change (cf. section 3.2.3). Similar to stars the stellar class was modeled using the highest number of mixture components as compared to galaxies or quasars. The covariance models determined by the algorithm remained the same when using only three classes.

Table 18: Mixture clustering with PCA results for $G = 17$ data using BP/RP only with stars and physical binaries merged, up to 6 mixture components per class, 40 principal components: Covariance models and number of mixture components per class.

	Galaxies	Quasar	Stellar
Covariance Model	VVV	VVV	VVV
#Mixtures	4	5	6

3.3.4 Radial basis function networks

Again we varied the number of basis functions between 1 and 250 as well as the regularisation parameter between 0 and 5 (table 19).

Table 19: RBF network results for $G = 17$ data using BP/RP only with stars and physical binaries merged: Total classification error [%].

Size	Decay					
	0	1	2	3	4	5
1	42.74	42.74	42.74	42.74	42.74	42.74
50	29.34	29.56	29.63	29.68	29.67	29.65
100	27.00	27.46	27.50	27.56	27.57	27.55
150	28.00	27.80	27.86	28.15	28.11	28.14
200	28.46	28.41	28.47	28.53	28.60	28.58
250	26.00	25.97	25.95	25.96	25.93	25.96

The confusion matrix for the best result (25.93%) using 250 basis functions and a weight decay of four is shown in table 20. The runtime for the training was 1675.0 sec (real time). The runtime for the evaluation was 252.8 sec.

Merging stars and physical binaries reduced the total classification error by 12.59% (cf. section 3.2.4). However the classification error is still significantly higher compared to neural networks (section 3.3.2) or mixture clustering (section 3.3.3). This is partly due to 43.01% of the galaxies which are now classified as stellar. This effect was also observed in the Adaboost.M1 experiments (section 3.3.1) and it increases the number of misclassifications considerably.

Table 20: RBF network results for $G = 17$ data using BP/RP only, with stars and physical binaries merged, $size = 250$, $decay = 4$: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	53.82	3.17	43.01
Quasar	5.94	78.42	15.64
Stellar	15.06	3.50	81.44

3.4 BP/RP and astrometry

In these experiments we included the features for parallax and proper motion. We expected that those features will lower the total classification error compared to the results in section 3.2.

3.4.1 Adaboost.M1-tests with adabag

For boosting the Adaboost.M1-implementation of the `adabag` R-library was used. Table 21 shows the total classification error for different iterations of the EM algorithm.

Table 21: Adaboost.M1 results for $G = 17$ data using BP/RP and astrometry: Total classification error and runtimes in real time.

Iterations	Total Error	RT Train	RT Eval
	%	sec	sec
1	23.29	20.1	1.2
100	13.42	1 591.2	87.2
200	13.07	3 275.9	214.8
300	13.23	4 939.3	380.8
400	13.09	6 586.4	603.1
500	13.03	8 183.3	879.0
600	13.05	10 256.9	1 207.6

The lowest total classification error (13.03%) was obtained for 500 iterations. For this result table 22 shows the confusion matrix. Adding the parallax and proper motion features had a considerable impact on the total classification error which was reduced by 24.78% (cf. section

3.2.1). Especially the classification of galaxies improved which is near perfect when using parallaxes and astrometry. Furthermore the confusion matrix proved to be relatively stable for all experiments with more than 100 iterations.

Table 22: Adaboost.M1 results for $G = 17$ data using BP/RP and astrometry: 500 iterations: Confusion matrix in percent.

True classes	Galaxy	Phys. Binary	Quasar	Star
Galaxy	99.50	0.00	0.50	0.00
Phys. Binary	0.00	75.73	0.00	24.27
Quasar	5.99	0.00	94.01	0.00
Star	0.00	21.53	0.00	78.47

3.4.2 Neural Networks

As in the experiments using four classes we varied the size of the hidden layer between 1 and 9 as well as the weight decay between 0 and 100 (table 23).

Table 23: Neural network results for $G = 17$ data using BP/RP and astrometry: Total classification error [%].

Size	Decay								
	0	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	100
1	30.91	31.05	77.08	77.75	77.65	78.03	77.75	99.83	97.68
2	24.01	14.17	13.06	12.81	13.07	14.24	16.61	21.00	35.26
3	15.65	18.76	14.37	12.33	12.10	11.23	11.74	16.21	33.50
4	12.86	14.89	14.04	10.20	9.41	8.61	10.00	16.99	33.73
5	13.92	16.58	13.58	10.06	9.18	8.35	10.00	15.94	31.18
6	12.81	10.56	12.33	10.66	8.91	8.11	9.69	17.03	33.56
7	16.16	14.58	12.26	11.23	9.49	7.87	8.76	15.91	31.61
8	13.60	13.01	11.14	10.67	9.94	7.46	8.96	16.16	31.83
9	11.72	13.74	11.51	11.59	11.09	8.03	8.60	15.30	31.27

The best result (7.46%) was obtained using $size = 8$ and $decay = 10^{-1}$. For this experiment the confusion matrix is shown in table 24. The runtime for the training was 366.2 sec (real time). The runtime for the predictions was 1.1 sec.

Compared to the Adaboost.M1 algorithm (section 3.4.1) neural networks again resulted in a lower classification error. Compared to the neural network experiments without astrometry the error rate was reduced by 1.98% (cf. section 3.2.2). Similar to the experiments without astrometry the most misclassifications took place between physical binaries and stars.

Table 24: Neural network results for $G = 17$ data using BP/RP and astrometry, $size = 8$, $decay = 10^{-1}$: Confusion matrix in percent.

True classes	Galaxy	Phys. Binary	Quasar	Star
Galaxy	99.95	0.00	0.05	0.00
Phys. Binary	0.05	86.29	0.00	13.66
Quasar	1.30	0.22	98.35	0.13
Star	0.14	14.26	0.14	85.46

3.4.3 Mixture clustering with PCA

The maximum number of mixture components per class was varied between 1 and 10. The number of principal components was also varied during the experiment between 1 (maximum data reduction) and 98 (no data reduction). Compared to the experiments without astrometry the maximum number of principal components is 98 (instead of 96) due to the two additional dimensions of the feature vector (parallax and proper motion). The results are shown in table 25.

Table 25: Mixture clustering with PCA results for $G = 17$ data using BP/RP and astrometry: Total classification error [%].

Mixtures	Number of PCs										
	1	10	20	30	40	50	60	70	80	90	98
$[1,1]$	61.02	27.78	22.86	20.01	19.27	19.66	19.73	20.04	19.90	20.00	19.73
$[1,2]$	58.11	19.10	14.17	13.15	13.57	13.20	14.15	14.61	14.91	15.44	15.42
$[1,3]$	56.81	16.79	11.63	11.25	11.53	13.23	13.73	11.97	12.66	13.52	16.67
$[1,4]$	56.97	15.07	11.71	9.56	11.16	12.70	13.38	11.58	15.24	13.46	16.05
$[1,5]$	56.30	13.58	9.94	9.55	9.49	10.69	10.91	11.44	12.22	13.07	15.40
$[1,6]$	56.28	12.43	9.99	9.55	9.38	10.65	10.91	11.45	12.22	–	15.36
$[1,7]$	56.28	12.41	10.00	9.55	9.38	10.65	10.91	–	–	–	–
$[1,8]$	56.28	12.35	10.00	9.52	9.38	10.65	10.91	–	–	–	–
$[1,9]$	56.28	12.35	10.00	9.51	9.38	10.65	10.91	–	–	–	–
$[1,10]$	56.28	12.27	10.00	9.51	9.38	10.65	–	–	–	–	–

The best result was a total error rate of 9.38% was obtained using 40 principal components and up to six mixture components per class. This result was reproduced when using up to seven, eight, nine or ten mixture components per class. The confusion matrix for 40 principal components and up to six mixture components per class is shown in table 26. The runtime for the training was 1 001.1 sec (real time). The runtime for the evaluation on the test set was 2.9 sec. Similar to neural networks (section 3.4.2) most misclassifications took place between physical binaries and stars.

Compared to the results without astrometry the total classification error was reduced by 1.55% (cf. section 3.2.3). Similar to those results the PCA and the subsequent dimensionality reduction were able to reduce the number of classification errors.

Table 26: Mixture clustering with PCA results for $G = 17$ data using BP/RP and astrometry, up to 6 mixture components per class, 40 principal components: Confusion matrix in percent.

True classes	Galaxies	Phys. Binary	Quasars	Stars
Galaxy	99.55	0.45	0.00	0.00
Phys. Binary	0.00	80.33	0.00	19.67
Quasars	0.00	0.04	99.55	0.40
Stars	0.00	17.38	0.00	82.62

For the same experiment table 27 shows the covariance models and the number of mixtures per class which were determined during training. Compared to the results without astrometry (section 3.2.3) the covariance models chosen by the algorithm did not change. For all classes except quasars the same number of mixture components was used. In the experiments with astrometry three instead of five mixture components were used.

Table 27: Mixture clustering with PCA results for $G = 17$ data using BP/RP and astrometry, up to 6 mixture components per class, 40 principal components: Covariance models and number of mixture components per class.

	Galaxies	Phys. Binary	Quasars	Stars
Covariance Model	VVV	VVV	VVV	VVV
#Mixtures	4	4	3	6

3.4.4 Radial basis function networks

Again we varied the number of basis functions between 1 and 250 as well as the decay parameter for regularisation between 0 and 5 (table 28).

Table 28: RBF network results for $G = 17$ data using BP/RP and astrometry: Total classification error [%].

Size	Decay					
	0	1	2	3	4	5
1	39.47	39.47	39.47	39.47	39.47	39.47
50	28.46	29.37	29.38	29.35	29.37	29.37
100	27.90	28.64	28.71	28.70	28.67	28.70
150	27.12	27.91	27.90	27.89	27.80	27.81
200	28.55	28.35	28.46	28.47	28.49	28.53
250	27.38	26.32	26.39	26.37	26.43	26.50

The confusion matrix for the best result (27.12%) using 150 basis functions and no decay is shown in table 29. The runtime for the training was 3 160.7 sec (real time). The runtime for the evaluation was 208.7 sec.

Similar to the Adaboost.M1 algorithm (section 3.4.1) the addition of astrometry had a significant impact on the total classification error of radial basis function networks (11.40%). However the error is still significantly higher than for neural networks (section 3.4.2 or mixture clustering (section 3.4.2). Similar to the experiments without astrometry the decay for the logistic regression does not have a significant impact on the total classification error (cf. section 3.2.4). Similar to Adaboost.M1 galaxies seem to profit most from astrometry. Here the total classification error was reduced by 23.99%.

Table 29: RBF network results for $G = 17$ data using BP/RP and astrometry with stars and physical binaries merged, $size = 150$, $decay = 0$: Confusion matrix in percent.

True classes	Galaxy	Phys. Binary	Quasar	Star
Galaxy	95.02	0.45	3.82	0.70
Phys. Binary	12.56	71.22	1.45	14.77
Quasar	9.74	0.45	84.05	5.76
Star	15.70	33.69	8.34	42.27

3.5 BP/RP and astrometry: Stars and physical binaries merged

As in the experiments involving BP/RP only we also investigated merging physical binaries and stars when using BP/RP together with astrometry.

3.5.1 Adaboost.M1-tests with adabag

The merging of stars and physical binaries had a significant effect when using BP/RP together with astrometry. As a difference to the experiments using four classes (section 3.4.1) the total error was reduced by 24%. this is a significant improvement compared to the experiments without astrometry (section 3.3.1) in which the merging of physical binaries and stars only led to a reduction of 3%. Compared to the Adaboost.M1 experiments using four classes the total classification error was reduced by 6.97%. Table 30 shows the total classification error for different iterations of the EM algorithm.

Table 30: Adaboost.M1 results for $G = 17$ data using BP/RP and astrometry with stars and physical binaries merged: Total classification error and runtimes in real time.

Iterations	Total Error	RT Train	RT Eval
	%	sec	sec
1	13.69	10.8	1.2
100	1.58	995.0	83.5
200	1.46	2 073.2	206.3
300	1.53	3 137.0	389.7
400	1.43	4 257.2	599.2
500	1.53	5 311.4	836.9
600	1.47	6 380.1	1 099.0

The lowest total classification error (1.43%) was obtained for 400 iterations. For this result table 31 shows the confusion matrix. As a difference to the experiments without astrometry (section 3.3.1) the misclassifications of galaxies did not increase which seems to be attributed to the astrometry. This seems to be the reason for the significant error reduction.

Table 31: Adaboost.M1 results for $G = 17$ data using BP/RP and astrometry with stars and physical binaries merged, 400 iterations: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	99.45	0.55	0.00
Quasar	4.87	95.13	0.00
Stellar	0.00	0.00	100.00

3.5.2 Neural Networks

Again we varied the size of the layer between 1 and 9 and the weight decay between 0 and 100 (table 32).

Table 32: Neural network results for $G = 17$ data using BP/RP and astrometry with stars and physical binaries merged: Total classification error [%].

Size	Decay								
	0	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	100
1	20.96	7.20	6.82	7.78	7.13	7.86	10.29	25.04	76.11
2	2.04	2.02	1.66	2.56	1.43	2.10	3.79	6.95	25.56
3	1.73	1.22	0.97	0.90	1.49	0.69	1.02	4.27	23.19
4	1.82	0.50	1.45	1.06	0.32	0.42	1.00	3.88	13.81
5	1.59	0.85	0.70	0.37	0.33	0.30	1.02	4.30	14.42
6	1.62	0.43	0.18	0.44	0.29	0.33	1.00	3.48	13.07
7	1.33	0.67	0.39	0.42	0.24	0.27	0.92	3.55	14.60
8	1.51	1.09	0.19	0.17	0.27	0.49	0.85	3.48	14.32
9	0.68	0.16	0.31	0.20	0.20	0.36	0.79	3.29	14.52

The best result (0.16%) was obtained using $size = 9$ and $decay = 10^{-5}$. Out of a total of 8360 objects only 13 were misclassified. The weight decay for the best result is lower compared to the other neural network experiments on $G = 17$ data. However this corresponds only to 26 errors which is not a big difference to the result for $size = 9$ and $decay = 10^{-1}$ (30 errors). Compared to the results using four classes (section 3.4.2) the error rate has been reduced by 7.30 % by merging stars and physical binaries. Compared to the results using no astrometry (section 3.3.2) the total classification error could be reduced by another 1.67%.

For the best result the confusion matrix is shown in table 33. The runtime for the training was 327.5 sec (real time). The runtime for the predictions was 0.1 sec. The classification for galaxies is perfect.

Table 33: Neural network results for $G = 17$ data using BP/RP only, $size = 9$, $decay = 10^{-5}$: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	100.00	0.00	0.00
Quasar	0.40	99.60	0.00
Stellar	0.05	0.05	99.90

3.5.3 Mixture clustering with PCA

The maximum number of mixture components per class was varied between 1 and 10. The number of principal components was also varied during the experiment between 1 (maximum data reduction) and 98 (no data reduction). The results are shown in table 34.

Table 34: Mixture clustering with PCA results for $G = 17$ data using BP/RP and astrometry with stars and physical binaries merged: Total classification error [%].

Mixtures	Number of PCs										
	1	10	20	30	40	50	60	70	80	90	98
$[1, 1]$	58.46	9.74	6.40	4.97	4.95	5.09	5.15	5.25	5.27	5.27	5.29
$[1, 2]$	49.84	3.18	0.85	0.94	2.00	2.22	3.00	3.39	3.36	3.47	3.19
$[1, 3]$	48.61	1.92	0.51	1.15	1.55	2.07	2.63	0.65	0.75	0.98	4.08
$[1, 4]$	48.32	1.21	0.41	0.12	1.66	1.60	2.01	0.47	3.55	0.97	3.50
$[1, 5]$	47.55	1.18	0.14	0.08	0.12	0.19	0.22	0.37	0.47	0.72	2.45
$[1, 6]$	47.91	0.94	0.22	0.08	0.10	0.17	0.22	0.37	0.47	–	2.26
$[1, 7]$	47.91	0.81	0.24	0.08	0.08	0.17	0.22	–	0.47	–	2.26
$[1, 8]$	48.70	0.70	0.24	0.10	0.08	0.17	0.22	–	0.47	–	2.70
$[1, 9]$	48.70	0.70	0.24	0.08	0.08	0.17	0.22	–	0.47	–	2.61
$[1, 10]$	48.70	0.59	0.24	0.08	0.08	0.17	0.22	–	–	–	–

The best result had a total error rate of 0.08% and was obtained using 30 or 40 principal components and a maximum of 5-10 mixture components per class. The confusion matrix for 30 principal components and up to five mixture components per class is shown in table 35. Again the dimension reduction in combination with the PCA was able to reduce the number of classification errors. The runtime for the training was 747.1 sec (real time). The runtime for the evaluation on the test set was 1.9 sec.

Table 35: Mixture clustering with PCA results for $G = 17$ data using BP/RP and astrometry with stars and physical binaries merged, up to five mixture components per class, 30 principal components: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	99.85	0.05	0.10
Quasar	0.00	99.82	0.18
Stellar	0.00	0.00	100.00

Compared to neural networks (section 3.5.2) mixture clustering achieved an even lower total classification error. Out of 8 360 objects only 7 were misclassified. Compared to the experiments using four classes (section 3.4.2) the error rate was reduced by 9.3%. This was mainly done by removing the misclassifications between the physical binaries and the stars which are now classified perfectly. Compared to experiments without astrometry (section 3.3.3) the error rate was reduced by another 0.72%.

For the same experiment table 36 shows the covariance models and the number of mixtures per class which were determined during training. As in the experiments using four classes (section 3.4.2) and in the experiments without astrometry (section 3.3.3) the "VVV" model was used for the covariance matrices in each of the classes. The numbers of mixture components are also similar.

Table 36: Mixture clustering with PCA results for $G = 17$ data using BP/RP and astrometry with stars and physical binaries merged, up to five mixture components per class, 30 principal components: Covariance models and number of mixture components per class.

	Galaxies	Quasar	Stellar
Covariance Model	VVV	VVV	VVV
#Mixtures	4	5	5

3.5.4 Radial basis function networks

Again we varied the number of basis functions between 1 and 250 as well as the regularisation parameter between 0 and 5 (table 37).

Table 37: RBF network results for $G = 17$ data using BP/RP and astrometry with stars and physical binaries merged: Total classification error [%].

Size	Decay					
	0	1	2	3	4	5
1	23.58	23.58	23.58	23.58	23.58	23.57
50	18.52	18.83	18.86	18.88	18.89	18.91
100	16.98	17.47	17.50	17.53	17.49	17.55
150	17.25	17.85	17.87	17.93	17.97	17.98
200	17.32	17.35	17.42	17.41	17.43	17.47
250	15.35	15.34	15.35	15.28	15.36	15.41

The confusion matrix for the best result (15.28%) using 250 basis functions and a decay of three is shown in table 38. The runtime for the training was 1 354.2 sec (real time). The runtime for the evaluation was 255.8 sec. Therefore radial basis function networks performed considerably worse than neural networks (section 3.5.2) or mixture clustering (section 3.5.3) on this data set which resulted in near perfect classifications.

Compared to the corresponding experiments without astrometry (section 3.3.4) the addition of the astrometry features resulted in a decrease of 10.65% of the total classification error. Compared to the experiments using four classes (section 3.4.4) the error rate was reduced by 11.84% because of merging physical binaries and stars. By this merger especially the number of misclassifications of stars was reduced considerably.

Table 38: RBF network results for $G = 17$ data using BP/RP and astrometry with stars and physical binaries merged, $size = 250$, $decay = 3$: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	92.91	1.11	5.99
Quasar	10.81	83.82	5.36
Stellar	14.00	4.73	81.27

4 Magnitude 20

4.1 Data sets

As for the $G = 17$ data the data set comprises 96 concatenated BP/RP bins as well as two features for parallaxes and proper motions (table 39). The data used are available under Subversion in the MPIA development directory (see Data file for $G = 20$ training; Data file for $G = 20$ evaluation in bibliography). Again the physical binaries used in the data had a brightness ratio of between zero and four and we used end-of-mission noise which was based on 80 transits.

Table 39: Data sets for $G = 20$ data.

Training set			
Galaxies	Physical Binaries	Quasars	Stars
1 934	2 073	2 156	2 162

Evaluation set			
Galaxies	Physical Binaries	Quasars	Stars
1 935	2 073	2 156	2 161

4.2 BP/RP only

In the tests in these sections only the BP/RP information (96 dimensions) of the data was used. The astrometry part (parallax and proper motions) was ignored.

4.2.1 Adaboost.M1-tests with adabag

As for the $G = 17$ data the Adaboost.M1-implementation of the `adabag` R-library R Project (2007); Cortés et al. (2007) was also used for the $G = 20$ data. Again we compared the total error for different numbers of iterations of the EM algorithm (table 40).

Table 40: Adaboost.M1 results for $G = 20$ data using BP/RP only: Total classification error [$\%$].

Iterations	Total Error
1	49.80
100	44.23
200	44.41
300	44.41
400	44.44
500	44.47
600	44.29

As the error rate does not seem to change significantly after 100 iterations we configured tests using 1-50 iterations. This time maximum CART tree depth values between 1 and 30 were additionally investigated (table 41).

Table 41: Adaboost.M1 results for $G = 20$ data using BP/RP only: Total classification error and runtimes in real time for EM iterations vs. CART tree depth.

Iterations	Max. Depth	Total Error	RT Train	RT Eval
		%	sec	sec
1	1	76.83	8.4	1.2
	10	45.88	41.6	1.2
	20	47.74	46.2	1.2
	30	46.87	48.5	1.2
10	1	83.62	62.5	8.2
	10	41.38	393.9	9.8
	20	42.34	531.0	9.3
	30	41.81	501.1	8.4
20	1	83.63	125.5	14.8
	10	41.63	798.7	24.6
	20	41.18	995.3	16.5
	30	41.13	1096.5	16.9
30	1	83.71	196.9	21.9
	10	40.97	1147.8	24.6
	20	40.84	1503.3	24.3
	30	41.29	1402.4	21.0
40	1	83.41	233.9	27.5
	10	40.19	1470.3	27.1
	20	40.82	1903.9	28.2
	30	41.21	1878.4	41.2
50	1	83.65	293.2	36.4
	10	40.90	1812.0	35.9
	20	40.89	2364.4	35.6
	30	40.77	2334.4	34.8

The lowest total classification error (40.19%) was obtained for 40 EM iterations and a CART tree depth of 10. For this result table 42 shows the confusion matrix. Therefore the variation of the CART tree depth resulted in another decrease of the classification error of about 4%. Compared to the experiments using $G = 17$ data (section 3.2.1) the total classification error was increased by 2.38%.

Table 42: Adaboost.M1 results for $G = 20$ data using BP/RP only: 100 iterations: Confusion matrix in percent.

True classes	Galaxy	Phys. Binary	Quasar	Star
Galaxy	66.30	27.49	1.34	4.86
Phys. Binary	21.85	71.15	0.05	6.95
Quasar	14.52	2.69	68.88	13.91
Star	25.68	38.92	1.34	34.06

4.2.2 Neural Networks

As for the $G = 17$ data we investigated neural networks. Again we varied the size of the hidden layer between 1 and 9 and chose values between 0 and 100 for the decay. The results are displayed in table 43.

Table 43: Neural network results for $G = 20$ data using BP/RP only: Total classification error [%].

Size	Decay								
	0	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	100
1	75.00	75.43	76.41	55.71	57.59	80.38	58.64	77.24	82.34
2	56.17	56.50	41.56	57.78	41.63	42.15	51.05	46.85	58.07
3	51.77	48.95	41.29	49.26	37.45	36.56	45.90	41.81	57.14
4	44.82	43.81	47.66	46.42	41.18	36.84	36.80	39.04	55.56
5	38.21	43.45	41.81	42.04	33.71	34.53	34.79	38.46	54.82
6	40.58	44.80	38.22	37.16	41.00	32.90	34.25	37.18	55.44
7	41.07	39.92	41.21	38.67	36.11	33.42	32.52	37.21	55.11
8	43.80	41.13	43.14	40.91	38.07	32.38	32.02	36.47	54.13
9	43.84	43.32	40.01	41.08	37.01	30.56	30.21	36.74	54.59

The best result (30.21%) was obtained using $size = 9$ and $decay = 1$. This corresponds to the results for the $G = 17$ data (section 3.2.2) for which $size = 9$ and $decay = 10^{-1}$ proved to be best. For $size = 9$ and $decay = 1$ the confusion matrix is shown in table 44. The runtime for the training was 480.1 sec (real time). The runtime for the predictions was 0.9 sec.

Therefore compared to the experiments using $G = 17$ data (section 3.2.2) the total classification error increased considerably by 20.77%. those misclassifications were not approximately evenly distributed among all classes. The total classification error for stars increased most

(31.67%) whereas the total classification error for quasars increased least (12.23%) for the best experiment. However as in the $G = 17$ experiments the classification error is again significantly lower than for the Adaboost.M1 algorithm (cf. section 4.2.1).

Table 44: Neural network results for $G = 20$ data using BP/RP only, $size = 9$, $decay = 1$: Confusion matrix in percent.

True classes	Galaxy	Phys. Binary	Quasar	Star
Galaxy	77.36	15.25	2.27	5.12
Phys. Binary	12.20	66.67	1.30	19.83
Quasar	5.10	1.44	84.51	8.95
Star	11.75	31.00	5.92	51.32

4.2.3 Mixture clustering with PCA

The maximum number of mixture components per class was varied between 1 and 10. The number of principal components was also varied during the experiment between 1 (maximum data reduction) and 96 (no data reduction). The three best parameter combinations resulted in a total classification error of 33.75% (up to 3 mixture components per class, 60 PCs), 33.79% (up to 7 mixture components per class, 50 PCs) and 33.85% (up to five mixture components per class, 50 PCs) respectively.

Table 45: Mixture clustering with PCA results for $G = 20$ data using BP/RP only: Total classification error [%].

Mixtures	Number of PCs									
	1	10	20	30	40	50	60	70	80	90	96
$[1,1]$	61.62	44.04	44.24	42.23	40.90	37.51	37.02	37.17	37.36	38.07	38.34
$[1,2]$	57.87	40.66	41.57	39.58	37.23	36.50	37.74	36.17	35.64	36.29	36.59
$[1,3]$	58.45	37.23	39.16	39.98	37.49	34.82	33.75	35.71	36.28	36.67	37.15
$[1,4]$	57.14	38.01	38.59	39.90	37.84	34.70	34.35	36.49	35.83	36.25	36.80
$[1,5]$	57.14	35.80	40.24	38.07	37.78	33.85	34.57	–	–	–	–
$[1,6]$	57.14	35.87	39.89	37.81	37.62	34.20	–	–	–	–	–
$[1,7]$	57.16	35.83	38.91	36.49	35.52	33.79	–	–	–	–	–
$[1,8]$	57.17	35.67	38.55	36.22	36.16	–	–	–	–	–	–
$[1,9]$	57.17	35.16	38.20	36.43	37.47	–	–	–	–	–	–
$[1,10]$	57.19	35.14	38.08	–	–	–	–	–	–	–	–

The best result had a total error rate of 33.75% and was obtained using 60 principal components and a maximum of three mixture components per class. The confusion matrix this result is shown in table 46. The runtime for the training was 875.1 sec (real time). The runtime for the evaluation on the test set was 3.3 sec. Again the dimension reduction in combination with the PCA was able to reduce the number of classification errors. For four mixture components per class using only 60 principal components instead of 96 reduced the error rate by 2.45%.

Similar to the experiments using $G = 17$ data (section 3.2.3) mixture clustering in combination with a PCA performs worse than neural networks. Moreover using higher numbers of mixture components (more than three) does not decrease the total classification error for $G = 20$ data. It is also interesting to note that the optimum number of principal components seems to be higher (50 to 60 for $G = 20$ instead of 20 to 30 for $G = 17$).

Table 46: Mixture clustering with PCA results for $G = 20$ data using BP/RP only, up to three mixture components per class, 60 principal components: Confusion matrix in percent.

True classes	Galaxy	Phys. Binary	Quasar	Star
Galaxy	73.33	16.69	1.71	8.27
Phys. Binary	15.97	61.60	0.39	22.05
Quasar	5.33	1.44	78.15	15.07
Star	14.16	31.75	1.62	52.48

For the experiment with the lowest total classification error table 47 shows the covariance models and the number of mixtures per class which were determined during training. As has already been discussed above the number of mixture components used to model each class is lower for the experiments on $G = 20$ data than for the experiments using $G = 17$ data (section 3.2.3).

Table 47: Mixture clustering with PCA results for $G = 20$ data using BP/RP only, up to three mixture components per class, 60 principal components: Covariance models and number of mixture components per class.

	Galaxies	Phys. Binary	Quasars	Stars
Covariance Model	EEE	VEV	EEE	VVV
#Mixtures	3	2	3	3

Moreover different covariance models were chosen (see Fraley & Raftery (2006)). “EEE” stands for covariance ellipsoids with equal volume, shape and orientation for every mixture component in a given class. “VEV” stands for covariance ellipsoids with equal shape but possibly different volume and orientation. “VVV” stands for covariance ellipsoids with arbitrary volume, shape and orientation. This means that compared to the most complex covariance model “VVV” which was used for all classes using the $G = 17$ simpler models were chosen. However this choice of models was not stable across the experiments in table 45. This result together with the higher optimum number of principal components seems to indicate that the $G = 20$ data could not be modeled as well as the $G = 17$ data by the mixture clustering algorithm.

4.2.4 Radial basis function networks

As for the $G = 17$ data we varied the number of basis functions between 1 and 250 as well as the regularisation parameter between 0 and 5 (table 48).

Table 48: RBF network results for $G = 20$ data using BP/RP only: Total classification error [%].

Size	Decay					
	0	1	2	3	4	5
1	49.93	49.97	50.01	50.02	50.03	50.02
50	34.68	34.91	34.88	34.93	34.92	34.92
100	36.02	36.30	36.35	36.38	36.35	36.41
150	37.77	38.55	38.51	38.55	38.53	38.53
200	39.83	39.82	39.89	39.98	40.05	40.05
250	40.23	39.87	39.94	39.95	40.02	40.02

The parameter ranges for the size and the weight decay of the RBF network classification were estimated using additional tests which are not displayed here. The lowest classification errors were obtained using 50 basis functions. We also evaluated the sizes 5, 10, 15 and 20 which did not improve the result for 50 basis functions.

The confusion matrix for the best result (34.68%) using 50 basis functions and no decay is shown in table 49. The runtime for the training was 2 587.3 sec (real time). The runtime for the evaluation was 170.0 sec.

Surprisingly the radial basis function network performs better on the $G = 20$ data than on the $G = 17$ data (cf. section 3.2.4). The total classification error was reduced by 3.84%. This seems mainly to be due to a decrease of the total classification error regarding stars which was reduced by 15.6%. This result proved to be mostly stable over all experiments in table 48. Compared to neural networks (section 4.2.1) radial basis function networks perform 4.47% worse regarding the total classification error. This is a considerably smaller difference than for the $G = 17$ data.

Table 49: RBF network results for $G = 20$ data using BP/RP only, $size = 50$, $decay = 0$: Confusion matrix in percent.

True classes	Galaxy	Phys. Binary	Quasar	Star
Galaxy	68.89	15.40	1.24	14.47
Phys. Binary	12.40	60.59	0.97	26.05
Quasar	5.33	1.58	83.02	10.07
Star	13.84	30.54	6.62	49.01

4.3 BP/RP only: Stars and physical binaries merged

As for the $G = 17$ data the classes for stars and physical binaries were merged as objects of these type are frequently confused and increase the difficulty of properly fitting the data.

4.3.1 Adaboost.M1-tests with adabag

Table 50 shows the results for the Adaboost.M1 algorithm for different number of iterations of the EM-algorithm.

Table 50: Adaboost.M1 results for $G = 20$ data using BP/RP only with stars and physical binaries merged: Total classification error and runtimes in real time.

Iterations	Total Error	RT Train	RT Eval
	%	sec	sec
1	85.73	19.5	1.4
100	35.44	1373.6	85.3
200	35.24	2686.6	202.9
300	35.18	4066.0	390.0
400	35.44	5372.4	604.0
500	35.38	6800.3	1060.3
600	35.46	8986.0	1193.4

The best result (35.18%) was obtained using 300 iterations. For this result the confusion matrix is shown in table 51. Compared to the experiments using all four classes (section 4.2.1) the error rate was reduced by 5.01%. Compared to the experiments on $G = 17$ data (section 3.3.1) the total classification error was increased by 1.26% due to the increase in magnitude. Similar to the $G = 17$ results the stellar objects were nearly perfectly classified. Again nearly all of the galaxies were misclassified. This seems to be an indication that when using only three classes the Adaboost.M1 algorithm needs astrometry information to properly distinguish between galaxies and stellar objects.

Table 51: Adaboost.M1 results for $G = 20$ data using BP/RP only with stars and physical binaries merged, 300 iterations: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	0.26	0.00	99.74
Quasar	0.65	54.69	44.67
Stellar	0.12	0.40	99.48

4.3.2 Neural Networks

Again we varied the size of the hidden layer between 1 and 9 and chose values between 0 and 100 for the weight decay. The results are displayed in table 52.

Table 52: Neural network results for $G = 20$ data using BP/RP only with stars and physical binaries merged: Total classification error [%].

Size	Decay								
	0	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	100
1	86.50	88.13	85.21	88.25	86.98	87.59	87.15	87.54	79.92
2	37.15	30.03	43.03	37.35	25.93	36.16	36.25	29.56	43.20
3	36.40	38.71	32.20	36.79	25.37	29.54	24.42	26.26	42.38
4	31.92	34.62	31.93	24.52	23.14	23.51	22.65	24.95	43.04
5	29.91	35.05	33.18	32.04	20.97	19.92	21.03	24.12	42.89
6	30.62	37.06	27.08	25.80	25.86	19.69	19.51	24.71	43.09
7	26.88	26.63	24.34	25.19	23.30	20.96	18.93	22.97	42.38
8	29.05	24.46	25.11	27.69	22.76	20.52	18.53	23.11	42.75
9	25.66	28.13	25.26	25.84	24.94	19.92	16.77	22.57	42.51

The best result (16.77%) was obtained using $size = 9$ and $decay = 1$. This corresponds to the results for the experiments using four classes (section 4.2.2). By merging stars and physical binaries the error rate was reduced by 13.44%. For $size = 9$ and $decay = 1$ the confusion matrix is shown in table 53. The runtime for the training was 484.7 sec (real time). The runtime for the predictions was 1.0 sec.

Compared to the experiments on $G = 17$ data (section 3.3.2) the total classification error increased by 14.94%. This increase seems to be mainly due to the misclassifications of galaxies of which 29.08% were misclassified. Most of the galaxies were wrongly classified as stellar objects. This result was stable over most of the experiments in table 52.

Compared to the experiments on $G = 20$ data for four classes (section 4.2.2) the total classification error was reduced by 13.44%. Compared to the Adaboost.M1 results on the $G = 20$ data neural networks again proved to result in a significantly lower total classification error (cf. section 4.3.1).

Table 53: Neural network results for $G = 20$ data using BP/RP only, $size = 9$, $decay = 1$: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	69.41	2.53	28.06
Quasar	5.84	82.75	11.41
Stellar	7.58	2.62	89.80

4.3.3 Mixture clustering with PCA

The maximum number of mixture components per class was varied between 1 and 10. The number of principal components was also varied during the experiment between 1 (maximum data reduction) and 96 (no data reduction). The results are shown in table 54.

Table 54: Mixture clustering with PCA results for $G = 20$ data using BP/RP only with stars and physical binaries merged: Total classification error [%].

Mixtures	Number of PCs										
	1	10	20	30	40	50	60	70	80	90	96
$[1,1]$	60.14	38.86	38.94	36.14	33.91	28.76	28.88	28.96	28.70	28.41	27.95
$[1,2]$	48.92	33.30	33.60	30.51	27.58	20.86	20.19	21.47	21.93	26.21	25.75
$[1,3]$	48.10	27.51	28.16	26.09	24.95	20.17	22.02	23.41	23.27	27.42	26.85
$[1,4]$	48.16	25.97	26.39	24.29	24.94	20.21	22.62	23.89	26.71	27.12	26.33
$[1,5]$	48.07	25.63	25.93	23.99	24.24	20.34	23.88	–	–	–	–
$[1,6]$	48.07	23.71	25.67	23.58	24.26	20.48	25.21	–	–	–	–
$[1,7]$	47.50	22.63	25.29	23.03	23.54	20.55	–	–	–	–	–
$[1,8]$	47.50	22.76	25.37	23.47	23.61	24.14	–	–	–	–	–
$[1,9]$	47.50	22.25	25.37	23.47	23.62	23.30	–	–	–	–	–
$[1,10]$	49.11	22.23	25.19	23.47	28.40	22.68	–	–	–	–	–

The best result had a total error rate of 20.17% and was obtained using 50 principal components and a maximum of three mixture components per class. The confusion matrix this result is shown in table 55. For the same experiment table 56 shows the covariance models and the number of mixtures per class which were determined during training. The runtime for the training was 1044.7 sec (real time). The runtime for the evaluation on the test set was 2.4 sec.

Again the dimension reduction in combination with the PCA was able to reduce the number of classification errors. For 3 mixture components per class using only 40 principal components instead of 96 reduced the error rate by 6.68%.

When using 50 principal components it is striking that when using up to 8 mixture instead of seven mixture the error rate rises by 3.59%. However this might be attributed to variances in the evaluation set which result in fewer errors when using fewer mixture components per class. In the training set a different number of mixture components per class might have proven to be best. However generally the error rate is decreasing when increasing the maximum number of mixture components per class.

Table 55: Mixture clustering with PCA results for $G = 20$ data using BP/RP only with stars and physical binaries merged, up to three mixture components per class, 40 principal components: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	75.76	0.98	23.26
Quasar	10.20	76.44	13.36
Stellar	15.82	0.76	83.42

Compared to the $G = 17$ data (section 3.3.3) the total classification error increased considerably by 19.37%. This increase in error was mainly due to misclassifications between galaxies and stellar objects. Compared to the experiments using four classes (section 4.2.3) the total classification error decreased by 10.04% due to the merger of both classes. Moreover similar to those experiments more principal components and less mixtures have been used than for the $G = 17$ data (section 3.3.3). The choice of covariance models is also similar to the experiments using four classes (section 4.2.3).

Table 56: Mixture clustering with PCA results for $G = 20$ data using BP/RP only with stars and physical binaries merged, up to three mixture components per class, 40 principal components: Covariance models and number of mixture components per class.

	Galaxies	Quasar	Stellar
Covariance Model	EEE	VEV	VVV
#Mixtures	3	3	3

4.3.4 Radial basis function networks

Again we varied the number of basis functions between 1 and 250 as well as the regularisation parameter between 0 and 5 (table 57).

Table 57: RBF network results for $G = 20$ data using BP/RP only with stars and physical binaries merged: Total classification error [%].

Size	Decay					
	0	1	2	3	4	5
1	43.02	43.02	43.02	43.02	43.02	43.02
50	20.23	20.88	21.00	21.05	21.07	21.12
100	20.94	21.17	21.27	21.30	21.27	21.26
150	21.53	22.01	21.99	22.02	21.93	21.89
200	22.55	22.94	23.15	23.21	23.24	23.32
250	21.81	22.68	22.88	22.86	22.95	22.97

The confusion matrix for the best result (20.23%) using 50 basis functions and no decay is shown in table 58. The runtime for the training was 293.9 sec (real time). The runtime for the evaluation was 71.4 sec.

Similar to the experiments using four classes (section 4.2.4) the radial basis function network performs better on the $G = 20$ data than on the $G = 17$ data (cf. section 3.3.4). Compared to those experiments the total classification error was reduced by 5.7% compared to the $G = 20$ data. However this time the reduction was evenly distributed among all three classes. Compared to the experiments using four classes the total classification error was reduced by 14.45%. Similar to the experiments using four classes the radial basis function networks compare considerably better compared to the neural networks on the $G = 20$ data than on the $G = 17$ data.

Table 58: RBF network results for $G = 20$ data using BP/RP only, with stars and physical binaries merged, $size = 50$, $decay = 0$: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	60.83	1.09	38.09
Quasar	5.33	82.47	12.20
Stellar	9.31	3.64	87.06

4.4 BP/RP and astrometry

4.4.1 Adaboost.M1-tests with adabag

Again we compared the total error for different numbers of iterations of the EM algorithm (table 59).

Table 59: Adaboost.M1 results for $G = 20$ data using BP/RP and astrometry: Total classification error and runtimes in real time.

Iterations	Total Error	RT Train	RT Eval
	%	sec	sec
1	46.91	18.8	1.2
100	42.61	1 786.5	81.5
200	42.68	4 112.2	242.0
300	43.03	5 799.2	386.7
400	42.69	7 409.1	571.7
500	42.85	9 530.1	835.8
600	42.56	11 070.2	1 129.8

The lowest total classification error (42.56%) was obtained for 600 iterations. For this result table 60 shows the confusion matrix. Compared to the results for $G = 17$ data (section 3.4.1) the total classification error increased considerably by 27.16%. This is also due to the fact that for the $G = 20$ data astrometry does not result in the same reduction of the total classification error than for the $G = 17$ data.

Table 60: Adaboost.M1 results for $G = 20$ data using BP/RP and astrometry: 600 iterations: Confusion matrix in percent.

True classes	Galaxy	Phys. Binary	Quasar	Star
Galaxy	91.06	5.58	0.67	2.69
Phys. Binary	32.61	54.70	1.83	10.85
Quasar	35.07	3.11	55.75	6.08
Star	24.11	34.11	10.13	31.65

4.4.2 Neural Networks

As in the tests without astrometry we investigated classification by means of neural networks. Again we varied the size of the layer as well as the value of decay for the weights of the neural network (table 61).

Table 61: Neural network results for $G = 20$ data using BP/RP and astrometry: Total classification error [%].

Size	Decay								
	0	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	100
1	51.17	51.11	88.18	57.06	75.53	51.45	52.08	53.84	88.83
2	44.71	47.17	44.56	37.15	37.19	37.47	40.06	40.91	50.10
3	40.40	43.91	41.13	31.59	39.04	33.49	33.72	34.44	47.46
4	37.91	40.59	40.90	36.53	30.68	30.35	33.11	33.39	44.14
5	38.85	36.67	35.95	33.81	33.08	30.73	30.26	32.56	45.95
6	36.90	34.94	36.46	35.35	35.95	27.28	29.02	31.94	43.87
7	37.63	34.23	36.68	33.05	36.22	27.96	27.98	31.65	44.32
8	35.04	36.25	37.51	34.11	32.92	27.63	27.26	30.77	43.54
9	34.59	33.45	36.45	35.17	34.34	27.63	26.75	31.52	43.30

The best result (26.75%) was obtained using $size = 9$ and $decay = 1$. The same parameter values were used to produce the best result for the $G = 20$ experiments without astrometry (section 4.2.2). For this run the confusion matrix is shown in table 62. The first column contains the true values of the observations. The runtime for the training was 523.0 sec (real time). The runtime for the predictions was 0.9 sec.

Compared to the $G = 20$ experiments without astrometry the total classification error was reduced by 3.46%. A similar reduction in error has been observed for the $G = 17$ data (cf. section 3.4.2).

Table 62: Neural network results for $G = 20$ data using BP/RP and astrometry, $size = 9$, $decay = 1$: Confusion matrix in percent.

True classes	Galaxy	Phys. Binary	Quasar	Star
Galaxy	83.62	10.08	2.84	3.46
Phys. Binary	10.67	64.45	0.92	23.98
Quasar	5.84	0.79	88.96	4.41
Star	7.96	29.85	5.46	56.73

4.4.3 Mixture clustering with PCA

The maximum number of mixture components per class was varied between 1 and 10. The number of principal components was also varied during the experiment between 1 (maximum data reduction) and 98 (no data reduction). The results are shown in table 63.

Table 63: Mixture clustering with PCA results for $G = 20$ data using BP/RP and astrometry: Total classification error [%].

Mixtures	Number of PCs										
	1	10	20	30	40	50	60	70	80	90	98
$[1,1]$	62.17	40.49	38.89	36.67	35.83	33.96	33.77	34.03	34.38	34.62	34.77
$[1,2]$	58.67	37.99	36.16	34.74	34.03	32.28	30.89	32.59	32.48	32.89	32.91
$[1,3]$	57.66	34.49	33.84	35.24	33.78	30.94	30.22	32.74	32.77	32.95	33.07
$[1,4]$	57.44	34.75	35.98	34.43	32.37	30.55	32.17	32.71	–	–	–
$[1,5]$	57.44	33.62	34.92	32.91	32.11	30.14	31.98	–	–	–	–
$[1,6]$	57.44	34.08	34.97	32.38	31.53	30.31	–	–	–	–	–
$[1,7]$	57.44	34.02	33.92	31.23	30.49	29.27	–	–	–	–	–
$[1,8]$	57.32	32.72	33.89	31.17	32.54	–	–	–	–	–	–
$[1,9]$	57.32	32.70	34.07	–	–	–	–	–	–	–	–
$[1,10]$	–	32.74	33.50	–	–	–	–	–	–	–	–

The best result had a total error rate of 30.14% and was obtained using 50 principal components and a maximum of five mixture components per class. The confusion matrix this result is shown in table 64. The runtime for the training was 1 109.8 sec (real time). The runtime for the evaluation on the test set was 3.0 sec. Again the dimension reduction in combination with the PCA was able to reduce the number of classification errors, but the effect was not as significant as in the experiments without astrometry (section 4.2.3). For 3 mixture components per class using only 60 principal components instead of 98 reduced the error rate by 2.85%.

Table 64: Mixture clustering with PCA results for $G = 20$ data using BP/RP and astrometry, up to five mixture components per class, 50 principal components: Confusion matrix in percent.

True classes	Galaxy	Phys. Binary	Quasar	Star
Galaxy	84.39	10.59	0.88	4.13
Phys. Binary	11.87	67.00	0.29	20.84
Quasar	8.35	1.35	77.23	13.08
Star	9.02	37.53	1.20	52.24

Compared to the experiments without astrometry (section 4.2.3) the total classification error was reduced by 3.61%. Compared to the corresponding results for the $G = 17$ data (cf. section 3.4.2) the total classification error increased by 20.76%. A similar increase in error took place for the experiments without astrometry (section 4.2.3).

For the experiment with the lowest total classification error table 65 shows the covariance models and the number of mixtures per class which were determined during training. Similar to the results without astrometry (section 4.2.3) for some classes simpler models were preferred over the most complex "VVV" covariance model. Moreover galaxies and quasars were modelled by five instead of two mixture components.

Table 65: Mixture clustering with PCA results for $G = 20$ data using BP/RP and astrometry, up to five mixture components per class, 50 principal components: Covariance models and number of mixture components per class.

	Galaxies	Phys. Binary	Quasars	Stars
Covariance Model	EEE	VVV	EEE	VVV
#Mixtures	5	2	5	3

4.4.4 Radial basis function networks

For radial basis function networks we used the RWeka library. We varied the number of basis functions between 1 and 250 as well as the regularisation parameter between 0 and 5 (table 66).

Table 66: RBF network results for $G = 20$ data using BP/RP and astrometry: Total classification error [%].

Size	Decay					
	0	1	2	3	4	5
1	46.34	46.34	46.34	46.34	46.34	46.34
50	–	30.56	30.63	30.63	30.64	30.65
100	32.28	32.83	32.96	33.03	33.02	33.11
150	33.13	33.74	33.72	33.78	33.75	33.74
200	33.61	33.98	34.02	34.02	34.08	34.07
250	35.54	36.14	36.30	36.30	36.34	36.36

The parameter ranges for the size and the weight decay of the RBF network classification were estimated using additional tests which are not displayed here. The best values were obtained using 50 basis functions. We also evaluated the sizes 5, 10, 15 and 20 which did not improve the result for 50 basis functions.

The experiment for size 50 and decay 0 does not seem to terminate which seems to be due to a software issue or a numeric instability of the algorithm. This was reproduced in several runs. The confusion matrix for the best result (30.56%) using 50 basis functions and a decay of 1 is shown in table 67. The runtime for the training was 407.1 sec (real time). The runtime for the evaluation was 79.2 sec.

Table 67: RBF network results for $G = 20$ data using BP/RP and astrometry, $size = 50$, $decay = 1$: Confusion matrix in percent.

True classes	Galaxy	Phys. Binary	Quasar	Star
Galaxy	77.73	12.66	1.09	8.53
Phys. Binary	9.31	62.57	0.58	27.55
Quasar	6.22	1.39	83.91	8.49
Star	8.79	32.35	4.67	54.19

As for the experiments without astrometry (section 4.2.4) there was only a relatively small increase in the total classification error when comparing the $G = 20$ results to the $G = 17$ results (section 3.4.4). The error increased by 3.44%. Therefore the error for radial basis function networks is only 3.81% higher than for the neural networks (cf. section 4.4.2).

4.5 BP/RP and astrometry: Stars and physical binaries merged

4.5.1 Adaboost.M1-tests with adabag

The merging of stars and physical binaries using the Adaboost.M1 algorithm had a similar effect for the $G = 20$ data than for the $G = 17$ data. Compared to the classification with stars and physical binaries in separate classes (section 4.4.1) the error rate was reduced by 16%. The reduction for the $G = 17$ data was 24% (section 3.5.1)

Table 68: Adaboost.M1 results for $G = 20$ data using BP/RP and astrometry with stars and physical binaries merged: Total classification error and runtimes in real time.

Iterations	Total Error	RT Train	RT Eval
	%	sec	sec
1	34.91	13.6	1.1
100	26.69	1171.9	86.0
200	25.39	2411.6	208.2
300	26.14	3641.4	394.9
400	26.29	4867.5	608.0
500	26.34	7035.3	1020.6
600	26.28	7947.6	1141.1

The lowest total classification error was obtained for 200 iterations (25.39%). For this result table 69 shows the confusion matrix. Compared to the experiments without astrometry (section 4.3.1) the total classification error was reduced by 9.75%. However as a difference to the experiments using the $G = 17$ data (section 3.4.1) the reduction of the total classification error was considerably lower.

Table 69: Adaboost.M1 results for $G = 20$ data using BP/RP and astrometry with stars and physical binaries merged: 200 iterations: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	62.07	3.10	34.83
Quasar	23.19	67.76	9.04
Stellar	10.09	6.09	83.82

4.5.2 Neural Networks

As in the tests using all four classes we investigated classification by means of ANNs. Again we varied the size of the layer as well as the weight decay of the neural network (table 70).

Table 70: Neural network results for $G = 20$ data using BP/RP and astrometry with stars and physical binaries merged: Total classification error [%].

Size	Decay								
	0	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	100
1	32.73	32.65	32.72	32.68	36.29	32.92	37.09	82.79	82.91
2	19.92	30.62	19.98	20.10	20.22	23.38	26.34	23.71	33.59
3	24.77	27.10	30.08	30.75	18.73	18.04	17.51	19.88	33.18
4	24.48	23.96	21.09	20.14	17.25	16.46	16.50	19.78	32.54
5	23.93	22.45	25.65	19.38	18.71	15.48	16.43	19.06	33.39
6	20.78	21.44	21.03	20.78	17.63	16.58	14.61	18.22	32.24
7	19.35	21.87	21.26	20.79	18.26	16.30	13.86	17.11	31.77
8	21.24	20.82	19.84	20.87	19.04	14.81	13.06	17.38	32.28
9	20.18	20.47	20.90	21.05	19.29	15.39	13.27	16.83	31.72

The best result (13.06%) was obtained using $size = 8$ and $decay = 1$. This is consistent with the $G = 20$ tests with classification of four classes. For this experiment the confusion matrix is shown in table 71. The first column contains the true values of the observations. The runtime for the training was 465.9 sec (real time). The runtime for the predictions was 1.1 sec. Compared to the experiments without astrometry (section 4.3.2) the total classification error was reduced by another 3.71%.

Table 71: Neural network results for $G = 20$ data using BP/RP and astrometry with stars and physical binaries merged, $size = 8$, $decay = 1$: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	80.21	3.10	16.69
Quasar	6.17	86.55	7.28
Stellar	6.47	3.31	90.22

4.5.3 Mixture clustering with PCA

The maximum number of mixture components per class was varied between 1 and 10. The number of principal components was also varied during the experiment between 1 (maximum data reduction) and 98 (no data reduction). The results are shown in table 72.

Table 72: Mixture clustering with PCA results for $G = 20$ data using BP/RP and astrometry with stars and physical binaries merged: Total classification error [%].

Mixtures	Number of PCs										
	1	10	20	30	40	50	60	70	80	90	98
1	60.16	32.04	29.38	24.52	23.10	21.32	20.98	21.17	21.33	21.31	21.15
2	49.62	28.42	25.09	21.73	20.44	16.28	15.68	17.35	17.55	18.93	18.94
3	48.79	23.36	21.45	19.89	19.18	16.56	16.14	18.07	19.75	19.42	19.11
4	48.61	22.69	20.77	17.54	17.92	16.80	17.55	17.67	–	–	–
5	48.36	21.06	20.06	16.98	17.33	16.97	17.65	18.11	–	–	–
6	48.36	20.52	20.25	16.97	17.07	17.17	–	–	–	–	–
7	48.36	20.49	19.28	15.98	16.17	16.52	–	–	–	–	–
8	48.65	19.78	19.41	15.98	18.28	17.78	–	–	–	–	–
9	48.65	19.81	19.41	17.85	19.72	17.49	–	–	–	–	–
10	48.65	19.65	19.41	17.85	–	16.90	–	–	–	–	–

The best result had a total error rate of 15.98% and was obtained using 30 principal components and a maximum of 7 mixture components per class. The confusion matrix this result is shown in table 73. The same result was obtained using up to 8 mixture components per class.

The runtime for the training was 997.9 sec (real time). The runtime for the evaluation on the test set was 1.9 sec. Again the dimension reduction in combination with the PCA was able to reduce the number of classification errors. For 3 mixture components per class using only 60 principal components instead of 98 reduced the error rate by 2.97%.

Similar to the $G = 17$ data (section 3.5.3) the total classification error was reduced significantly (13.69%, cf. section 4.4.3) by merging stars and physical binaries. The addition of astrometry additionally reduced the total classification error by 3.71% (cf. section 4.3.4).

Table 73: Mixture clustering with PCA results for $G = 20$ data using BP/RP and astrometry with stars and physical binaries merged, up to 7 mixture components per class, 30 principal components: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	87.49	2.89	9.61
Quasar	9.79	84.28	5.94
Stellar	15.35	2.34	82.31

For the experiment with the lowest total classification error table 74 shows the covariance models and the number of mixtures per class which were determined during training. As for the other mixture clustering experiments using $G = 20$ data simpler covariance models were used instead of the most complex "VVV" model. Here the "EEE" model (covariance ellipsoids of equal volume, shape and orientation, see Fraley & Raftery (2006)) was used to model galaxies. However as a difference to the experiments without astrometry (section 4.3.4) more mixture components per class were used.

Table 74: Mixture clustering with PCA results for $G = 20$ data using BP/RP and astrometry with stars and physical binaries merged, up to 7 mixture components per class, 30 principal components: Covariance models and number of mixture components per class.

	Galaxies	Quasar	Stellar
Covariance Model	EEE	VVV	VVV
#Mixtures	7	4	4

4.5.4 Radial basis function networks

Again we varied the number of basis functions between 1 and 250 as well as the regularisation parameter between 0 and 5 (table 75).

Table 75: RBF network results for $G = 20$ data using BP/RP and astrometry with stars and physical binaries merged: Total classification error [%].

Size	Decay					
	0	1	2	3	4	5
1	39.47	39.47	39.47	39.47	39.47	39.48
50	14.59	14.92	14.89	14.99	15.05	15.12
100	15.74	16.13	16.16	16.13	16.16	16.18
150	16.50	17.14	17.20	17.18	17.19	17.20
200	17.14	18.16	18.29	18.40	18.45	18.52
250	18.22	18.71	18.94	19.09	19.20	19.26

The confusion matrix for the best result (14.59%) using 50 basis functions and no decay is shown in table 76. The runtime for the training was 367.7 sec (real time). The runtime for the evaluation was 72.7 sec. Similar to the experiments without astrometry (section 4.3.4) the increase in magnitude results in a lower total classification error than on the $G = 17$ data. This suggests that the radial basis function network fits the $G = 20$ data considerably better than the $G = 17$ data. Moreover the radial basis function network performs better than mixture clustering (section 4.5.3) and only slightly worse than neural networks (section 4.5.2).

Table 76: RBF network results for $G = 20$ data using BP/RP and astrometry with stars and physical binaries merged, $size = 50$, $decay = 0$: Confusion matrix in percent.

True classes	Galaxy	Quasar	Stellar
Galaxy	75.97	0.62	23.41
Quasar	7.05	84.14	8.81
Stellar	7.20	2.43	90.36

5 Summary and Conclusions

In the experiments described in this report we investigated four algorithms for discrete source classification (DSC) which were applied to the Gaia cycle 2A data. We investigated $G = 17$ data as well as $G = 20$ data. Moreover we conducted tests with astrometry (BP/RP+Astrometry) and without astrometry (BP/RP) to investigate its impact on the DSC. We also investigated which impact the merging of stars and physical binaries had on the classification. This was due to the fact that for the given range of brightness ratio for physical binaries (zero to four) stars and physical binaries were generally hard to distinguish. The resulting misclassifications had a considerable impact on the overall perfomance of the classification algorithms. The merger resulted in a three class classification problem (galaxies, quasars, stellar objects) instead of the given four class classification problem (galaxies, physical binaries, quasars, stars). The number of classes used is denoted with K. This resulted in eight different classification tasks for each of the four algorithms.

In the experiments we used approximately the same number of sources in each of the four available classes in the training and test sets. Therefore we did not make use of prior probabilities which were approximately equal in each of the four classes. Real data however usually contains a considerably larger number of stars than galaxies or quasars. Therefore in this case prior probabilities will have to be taken into account.

As classification algorithms we investigated boosting using the Adaboost.M1 algorithm in connection with classification and regression trees. Moreover we investigated neural networks (ANN) using a single hidden layer. We also investigated mixture clustering (Mclust) which uses a set of Gaussian mixture components to model each class. In connection with mixture clustering we also used a principal component analysis for dimensionality reduction in order to make classification for larger numbers of mixture components possible. Furthermore radial basis function (RBF) networks using normalised Gaussians were used.

Table 77: Total classification error [%] for DSC algorithm.

	$G = 17$				$G = 20$			
	BP/RP		BP/RP+Astrometry		BP/RP		BP/RP+Astrometry	
	$K = 4$	$K = 3$	$K = 4$	$K = 3$	$K = 4$	$K = 3$	$K = 4$	$K = 3$
SVM	10.36	2.51	8.26	0.59	33.77	20.08	28.80	14.85
Boosting	37.81	33.93	13.03	1.46	44.23	35.18	42.61	25.39
ANN	$\underline{9.44}$	1.83	$\underline{7.46}$	0.16	$\underline{30.21}$	$\underline{16.77}$	$\underline{26.75}$	$\underline{13.06}$
Mclust	10.93	$\underline{0.80}$	9.38	$\underline{0.08}$	33.75	20.17	30.14	15.98
RBF	38.52	23.95	27.12	15.28	34.68	20.23	30.56	14.59

Table 77 shows the best results of each algorithm with respect to the eight classification tasks regarding the total classification error. For comparison the baseline results for the current DSC algorithm which is based on support vector machines (SVM) were added. The lowest total classification error for each classification task is underlined.

The results show that for the given algorithms ANN seems to perform best over all tasks. However when stars and physical binaries were merged on the $G = 17$ data the mixture clustering in combination with a PCA seems to perform slightly better. However it is possible that a PCA in connection with ANN would also result in another reduction of the ANN classification error. In these experiments we investigated the use of PCA only in connection with Mclust. It is surprising that on the $G = 20$ data RBF networks seem to work even better than on the $G = 17$ data. This is also reflected in smaller numbers of basis functions on the $G = 20$ data. Compared to the baseline SVM approach ANN perform better in each condition. The differences between SVM and Mclust are marginal except when using BP/RP only and three classes on the $G = 17$ data. Here SVM results in a 1.71% lower total classification error than Mclust.

It should be noted that for the Adaboost.M1 algorithm using $G = 20$ data and $K = 4$ classes we did not use the best result from table 41 (40.19%) but the best result from table 40 (44.23%). This is due to the fact that in table 41 we conducted an additional optimization of the CART tree depth. This tree depth optimization was not conducted in any other Adaboost.M1 experiments due to the relatively poor overall performance of the Adaboost.M1 algorithm compared to the other algorithms. The results suggest however that similar improvements for the Adaboost.M1 algorithm might also be possible in the other experiments. However the difference between the tree depth optimized Adaboost.M1 algorithm and the ANN is still 9.98% (table 77). Therefore we considered it unlikely that the Adaboost.M1 algorithm with optimized tree depth would outperform the ANN.

Table 78: Training runtimes (real time in sec) for DSC algorithm.

	$G = 17$				$G = 20$			
	BP/RP		BP/RP+Astrometry		BP/RP		BP/RP+Astrometry	
	$K = 4$	$K = 3$	$K = 4$	$K = 3$	$K = 4$	$K = 3$	$K = 4$	$K = 3$
SVM	98.48	48.85	77.12	31.45	166.88	144.92	151.62	108.75
Boosting	1 592.8	2 349.3	8 183.3	4 257.2	1 470.3	4 066.0	11 070.2	2 411.6
ANN	508.2	345.1	366.2	327.5	480.1	484.7	523.0	465.9
Mclust	736.6	1 302.6	1 001.1	747.1	875.1	1 044.7	1 109.8	997.9
RBF	1 257.9	1 675.0	3 160.7	1 354.2	2 587.3	293.9	407.1	367.7

Concerning Mclust we observed differences in the choice of the covariance models for the mixture components for each class. On $G = 17$ data more variable models regarding the volume, shape and orientation of the covariance ellipsoid were preferred whereas on $G = 20$ data simpler models were additionally used. Regarding the PCA we also observed differences between mixture clustering experiments on $G = 17$ and $G = 20$ data. On $G = 20$ data there seems to be a tendency of using more principal components. Moreover in some experiments lower numbers of mixture components were used to model each class on the $G = 20$ data.

On a more general level adding astrometry resulted in a reduction of the total classification error in nearly every experiment which stresses the importance of the astrometry features for the DSC algorithm. Even larger reductions of the total classification error could be observed when physical binaries and stars were merged ($K = 3$). This is due to the frequent misclassifications between physical binaries and stars which could be observed in all of the experiments.

An exception was the Boosting algorithm. Here nearly all of the galaxies were misclassified as stellar when using $K = 3$ and BP/RP only. This result was reproduced on $G = 17$ as well as on $G = 20$ data. However when using astrometry no corresponding effect took place. As a result the reduction of the total classification error when using $K = 3$ instead of $K = 4$ is lower for Boosting than for any other algorithm when using BP/RP only. This might indicate that astrometry information is necessary to properly classify the galaxies by means of the Boosting algorithm. However these results might also be due to numeric instabilities of the Boosting algorithm. Therefore the interpretation of these results requires further analysis.

Table 78 contains the runtimes for the training of the experiments from table 77. Here ANN and RBF networks were generally fastest during training. RBF networks were faster on the $G = 20$ data than on the $G = 17$ data because lower numbers of basis functions were used. However the training of ANN and RBF networks was still considerably slower than the baseline SVM training in each condition. It is important to note that the times measured are real times which depend on the processor load. However during the experiments the influence of processor load on the real time was observed to be generally small. Therefore the real time should provide a reasonable measure for the efficiency of the algorithms.

Table 79: Evaluation runtimes (real time in sec) for DSC algorithm.

	$G = 17$				$G = 20$			
	BP/RP		BP/RP+Astrometry		BP/RP		BP/RP+Astrometry	
	$K = 4$	$K = 3$	$K = 4$	$K = 3$	$K = 4$	$K = 3$	$K = 4$	$K = 3$
SVM	39.55	18.82	33.05	13.23	68.15	51.67	64.63	42.92
Boosting	79.9	191.15	879.0	599.2	27.1	390.0	1 129.8	208.2
ANN	1.1	1.1	1.1	0.1	0.9	1.0	0.9	1.1
Mclust	2.4	2.2	2.9	1.9	3.3	2.4	3.0	1.9
RBF	142.3	252.8	208.7	255.8	170.0	71.4	79.2	72.7

Table 79 shows the runtimes of the experiments on the evaluation sets. Here ANN and Mclust proved to be fastest. However as a difference to the other algorithms mixture clustering took advantage of the principal component analysis which was applied as a preprocessing step to allow experiments with larger numbers of mixture components. As a difference to the training runtimes the evaluation of SVM took longer than the evaluation of ANN or Mclust.

6 Next Steps

The results indicate that the ANNs are suitable for fitting the cycle 2A data. Therefore the use of ANNs will be investigated further in the future. In connection with the Mclust algorithm it also became apparent that a dimension reduction (in this case the PCA) has a positive impact on the total classification error. Therefore we will investigate different methods of preprocessing the Gaia data. Apart from the linear approach the PCA is taking we will also take nonlinear preprocessing approaches like principle curves into account.

7 References

Cortés E., Gámez M., García N., August 2007, International Advances in Economic Research, 13, 301

Data file for $G = 17$ evaluation, 2007, `http://gaia.esac.esa.int/dpacsvn/DPAC/CU8/MPIA/dsc_cycle2_experiments/m17WithNoiseNewComp2.txt`

Data file for $G = 17$ training, 2007, `http://gaia.esac.esa.int/dpacsvn/DPAC/CU8/MPIA/dsc_cycle2_experiments/m17WithNoiseNewComp1.txt`

Data file for $G = 20$ evaluation, 2007, `http://gaia.esac.esa.int/dpacsvn/DPAC/CU8/MPIA/dsc_cycle2_experiments/m20WithNoiseNewComp2.txt`

Data file for $G = 20$ training, 2007, `http://gaia.esac.esa.int/dpacsvn/DPAC/CU8/MPIA/dsc_cycle2_experiments/m20WithNoiseNewComp1.txt`

Fraley C., Raftery A.E., 2006, MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering, Tech. rep., Department of Statistics, University of Washington, http://www.stat.washington.edu/fraley/tr504.pdf

R Project (2007), 2007, The R Project for Statistical Computing, http://www.r-project.org/

Venables W.N., Ripley B.D., 2002, Modern Applied Statistics with S, Springer