MIREX 2011 AUDIO TAG CLASSIFICATION USING WEIGHTED-VOTE NEAREST NEIGHBOR CLASSIFICATION

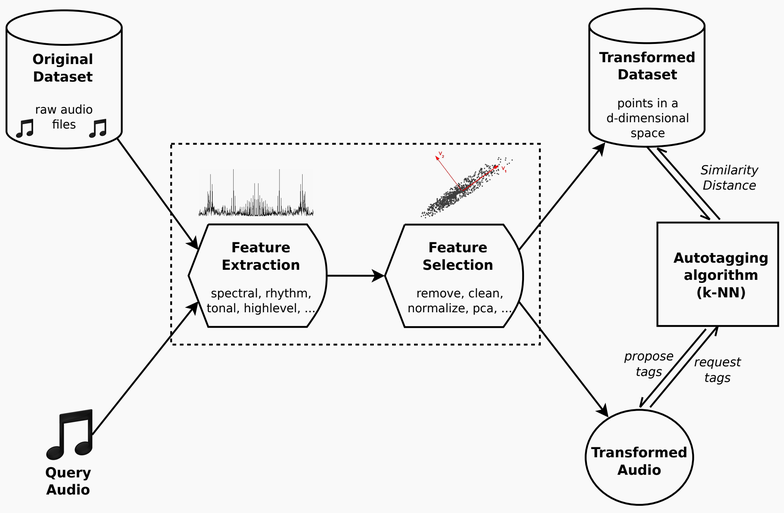
Mohamed Sordo[1], Òscar Celma[2], Dmitry Bogdanov[1]

[1] Music Technology Group, Univesritat Pompeu Fabra *{name.surname@upf.edu}*

[2] Gracenote *{ocelma@gmail.com}*

Method

General Framework



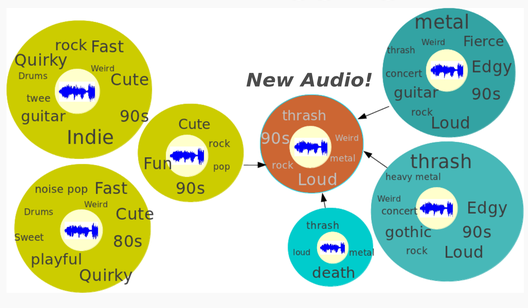
Feature Extraction

Low level	Rhythm/Tempo	Tonal	High level
Average loudness	Beat position	Chords	Genres
Bark bands	Beat loudness	Key	Moods
Dissonance	BPM	HPCP	Gender
HFC	Onset	Tuning	Speech/music
13–MFCC			Live/studio
Pitch			...
Silence			
Zero–crossing rate			
Spectral features (centroid, rollof, kurtosis,...)			

Table 1: Summary list of audio features used by our algorithm.

The audio features are captured on a short-time frame-by-frame basis, and then averaged over the whole audio excerpt. We take the means, variances and their corresponding deltas.

Autotagging



Feature Selection + Distance Measures

* **SC1**: PCA (75% covered variance) + Euclidean distance (EUC)

* **SBC1**: Linear combination (PCA-EUC, Tempo-based distance, Kullback Leibler divergence with 1G MFCC, Semantic classifier-based distance)

Affinity Ranking

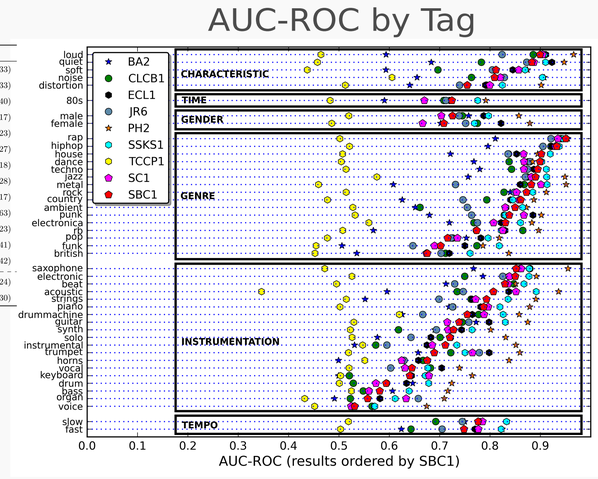
MajorMiner Dataset

Overall by Fold

Algorithm	AUC-ROC		P@3		P@6		P@9		P@12		P@15	
BA1	0.7798	(0.0039)	0.4411	(0.0061)	0.3720	(0.0107)	0.2976	(0.0041)	0.2378	(0.0043)	0.1954	(0.0033)
BA2	0.7801	(0.0017)	0.4342	(0.0123)	0.3737	(0.0070)	0.2998	(0.0035)	0.2387	(0.0032)	0.1963	(0.0033)
BA3	0.7793	(0.0058)	0.4396	(0.0086)	0.3722	(0.0092)	0.2969	(0.0053)	0.2382	(0.0044)	0.1966	(0.0040)
CCL1	0.7900	(0.0006)	0.2525	(0.0096)	0.2596	(0.0017)	0.2446	(0.0036)	0.2225	(0.0032)	0.2005	(0.0017)
CLCB1	0.8120	(0.0033)	0.3316	(0.0057)	0.2967	(0.0005)	0.2641	(0.0020)	0.2375	(0.0017)	0.2125	(0.0023)
ECL1	0.7980	(0.0019)	0.2627	(0.0080)	0.2619	(0.0003)	0.2458	(0.0052)	0.2233	(0.0039)	0.2033	(0.0027)
JR4	0.8533	(0.0007)	0.4895	(0.0063)	0.3905	(0.0046)	0.3122	(0.0043)	0.2613	(0.0028)	0.2240	(0.0018)
JR5	0.8502	(0.0018)	0.4826	(0.0082)	0.3878	(0.0045)	0.3104	(0.0035)	0.2598	(0.0047)	0.2220	(0.0028)
JR6	0.8290	(0.0016)	0.5049	(0.0005)	0.3864	(0.0039)	0.3076	(0.0024)	0.2540	(0.0021)	0.2157	(0.0017)
PH2	**0.9094**	(0.0261)	**0.5902**	(0.0539)	**0.4473**	(0.0338)	**0.3522**	(0.0202)	**0.2883**	(0.0119)	**0.2439**	(0.0063)
SSKS1	0.8917	(0.0026)	0.5510	(0.0104)	0.4286	(0.0054)	0.3410	(0.0062)	0.2820	(0.0037)	0.2391	(0.0023)
TCCP1	0.7942	(0.0040)	0.3783	(0.0123)	0.3100	(0.0030)	0.2688	(0.0033)	0.2336	(0.0049)	0.2013	(0.0041)
TCCP2	0.7937	(0.0042)	0.3775	(0.0114)	0.3094	(0.0026)	0.2682	(0.0036)	0.2335	(0.0050)	0.2013	(0.0042)
SBC1	0.8725	(0.0031)	0.5321	(0.0146)	0.4107	(0.0056)	0.3266	(0.0039)	0.2729	(0.0025)	0.2322	(0.0024)
SC1	0.8704	(0.0029)	0.5201	(0.0140)	0.4090	(0.0092)	0.3234	(0.0051)	0.2696	(0.0039)	0.2296	(0.0030)

Table 2: Comparative results (means and standard deviations) for Precision-At-N, using the Major Miner Dataset. The best results are indicated in bold.

AUC-ROC by Tag



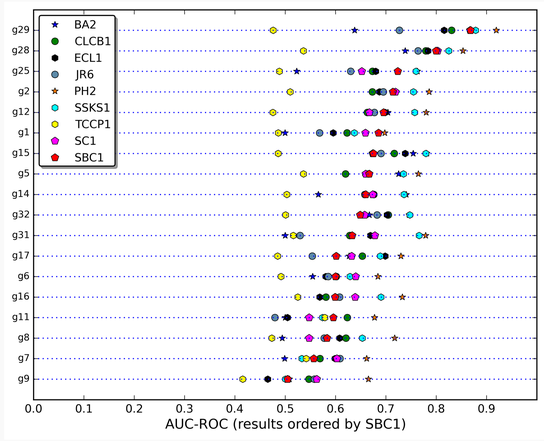
Mood Tag Dataset

Overall by Fold

Algorithm	AUC-ROC		P@3		P@6		P@9		P@12		P@15	
BA1	0.7890	(0.0106)	0.3117	(0.0050)	0.2421	(0.0043)	0.1714	(0.0033)	0.1376	(0.0005)	0.1190	(0.0018)
BA2	0.7769	(0.0103)	0.2904	(0.0091)	0.2418	(0.0046)	0.1702	(0.0049)	0.1372	(0.0034)	0.1181	(0.0026)
BA3	0.7883	(0.0144)	0.3152	(0.0160)	0.2441	(0.0059)	0.1717	(0.0036)	0.1383	(0.0012)	0.1196	(0.0016)
CCL1	0.6945	(0.0101)	0.1896	(0.0035)	0.1879	(0.0026)	0.1607	(0.0026)	0.1370	(0.0017)	0.1189	(0.0010)
CLCB1	0.7444	(0.0096)	0.2417	(0.0136)	0.2087	(0.0062)	0.1725	(0.0035)	0.1441	(0.0019)	0.1223	(0.0013)
ECL1	0.6921	(0.0089)	0.1780	(0.0109)	0.1814	(0.0024)	0.1604	(0.0019)	0.1383	(0.0004)	0.1195	(0.0009)
JR4	0.8406	(0.0009)	0.3739	(0.0015)	0.2528	(0.0029)	0.1877	(0.0021)	0.1483	(0.0016)	0.1220	(0.0014)
JR5	0.8434	(0.0009)	0.3752	(0.0009)	0.2555	(0.0050)	0.1889	(0.0012)	0.1481	(0.0013)	0.1220	(0.0010)
JR6	0.7994	(0.0058)	0.3684	(0.0018)	0.2406	(0.0020)	0.1768	(0.0007)	0.1394	(0.0000)	0.1169	(0.0005)
PH2	**0.8739**	(0.0130)	0.4033	(0.0131)	**0.2646**	(0.0021)	**0.1941**	(0.0005)	**0.1517**	(0.0009)	**0.1236**	(0.0009)
SSKS1	0.8654	(0.0018)	**0.4124**	(0.0051)	0.2629	(0.0029)	0.1902	(0.0017)	0.1489	(0.0010)	0.1223	(0.0012)
TCCP1	0.8262	(0.0056)	0.3467	(0.0074)	0.2515	(0.0060)	0.1860	(0.0037)	0.1485	(0.0019)	0.1225	(0.0015)
TCCP2	0.8262	(0.0059)	0.3458	(0.0076)	0.2515	(0.0060)	0.1858	(0.0034)	0.1487	(0.0016)	0.1223	(0.0016)
SBC1	0.8448	(0.0007)	0.3876	(0.0040)	0.2573	(0.0015)	0.1895	(0.0023)	0.1498	(0.0017)	0.1231	(0.0014)
SC1	0.8455	(0.0008)	0.3887	(0.0006)	0.2582	(0.0028)	0.1901	(0.0017)	0.1499	(0.0019)	0.1231	(0.0013)

Table 4: Comparative results (means and standard deviations) for Precision-At-N, using the Mirex'09 Mood Tag Dataset. The best results are indicated in bold.

AUC-ROC by Tag



Significance tests

* **AUC-ROC by Fold**: No statistical significance between all the presented algorithms.

* **AUC-ROC by Tag**: PH2>SBC1 is statistically significant. PH2>SC1 significant in MajorMiner only. SSKS1 is not statistically significant in both datasets. SBC1>(JR4-5, BAx, TCCPx) and SC1>(JR5, BAx, TCCPx) are statistically signficant in MajorMiner. SC1>SBC1 not stat. significant.

* **Precision at N**: No statistical significance between all the presented algorithms.

Classification

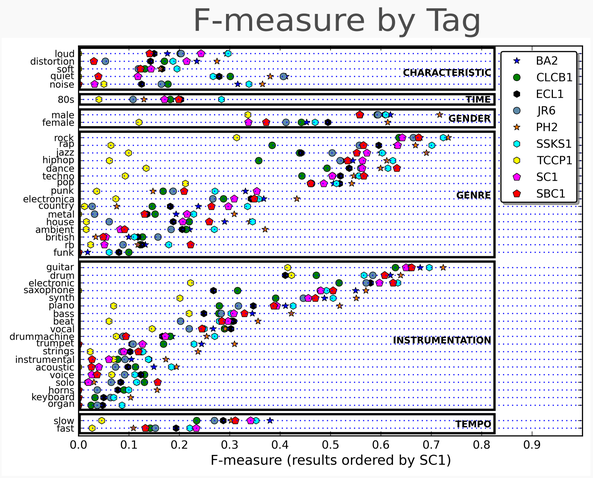
MajorMiner Dataset

Overall by Fold

Algorithm	Accuracy		Neg. Ex. Accuracy		Precision		Recall		F–measure	
BA1	0.8656	(0.0037)	0.8861	(0.0039)	0.3629	(0.0114)	0.6581	(0.0057)	0.4677	(0.0105)
BA2	0.8653	(0.0038)	0.8854	(0.0045)	0.3629	(0.0097)	**0.6621**	(0.0028)	0.4687	(0.0078)
BA3	0.8640	(0.0037)	0.8841	(0.0053)	0.3598	(0.0080)	0.6604	(0.0135)	0.4656	(0.0057)
CCL1	0.7936	(0.0011)	0.8138	(0.0008)	0.2374	(0.0039)	0.5880	(0.0059)	0.3382	(0.0048)
CLCB1	0.8015	(0.0008)	0.8178	(0.0007)	0.2560	(0.0017)	0.6362	(0.0020)	0.3651	(0.0017)
ECL1	0.7942	(0.0020)	0.8142	(0.0012)	0.2388	(0.0049)	0.5915	(0.0112)	0.3403	(0.0067)
JR4	0.8565	(0.0248)	0.8969	(0.0219)	0.2995	(0.0715)	0.4427	(0.1043)	0.3558	(0.0833)
JR5	0.8500	(0.0232)	0.8896	(0.0169)	0.2872	(0.0723)	0.4465	(0.0937)	0.3493	(0.0821)
JR6	0.8800	(0.0005)	0.9101	(0.0005)	0.3863	(0.0038)	0.5742	(0.0023)	0.4619	(0.0033)
PH2	0.9170	(0.0127)	0.9478	(0.0120)	0.5298	(0.0670)	0.5985	(0.0711)	**0.5601**	(0.0602)
SSKS1	**0.9105**	(0.0018)	**0.9494**	(0.0013)	0.4848	(0.0109)	0.5012	(0.0083)	0.4929	(0.0095)
TCCP1	0.8608	(0.0014)	0.9230	(0.0004)	0.2064	(0.0034)	0.2100	(0.0052)	0.2082	(0.0042)
TCCP2	0.8618	(0.0021)	0.9235	(0.0006)	0.2116	(0.0054)	0.2154	(0.0085)	0.2135	(0.0068)
SBC1	0.8731	(0.0018)	0.8940	(0.0016)	0.3804	(0.0050)	**0.6613**	(0.0038)	0.4830	(0.0050)
SC1	0.8712	(0.0028)	0.8925	(0.0032)	0.3749	(0.0087)	0.6545	(0.0095)	0.4767	(0.0085)

Table 3: Comparative results (means and standard deviations) for Binary Relevance, using the Major Miner Dataset. The best results are indicated in bold.

F-measure by Tag



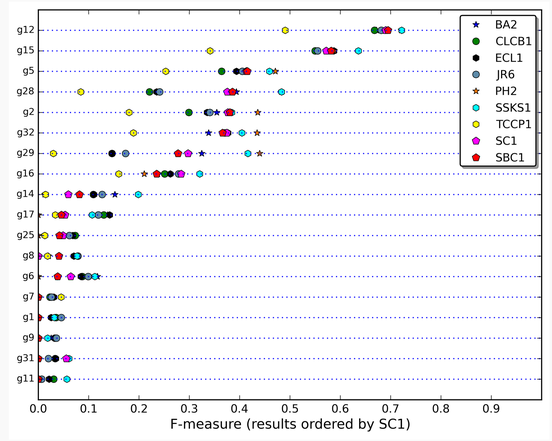
Mood Tag Dataset

Overall by Fold

Algorithm	Accuracy		Neg. Ex. Accuracy		Precision		Recall		F–measure	
BA1	0.7420	(0.0117)	0.7363	(0.0159)	0.2585	(0.0075)	0.7915	(0.0263)	0.3895	(0.0058)
BA2	0.7526	(0.0140)	0.7490	(0.0181)	0.2663	(0.0090)	0.7837	(0.0229)	0.3973	(0.0080)
BA3	0.7392	(0.0068)	0.7327	(0.0101)	0.2566	(0.0051)	0.7952	(0.0268)	0.3878	(0.0056)
CCL1	0.5090	(0.0021)	0.4741	(0.0012)	0.1517	(0.0016)	0.8109	(0.0122)	0.2555	(0.0027)
CLCB1	0.5204	(0.0021)	0.4804	(0.0012)	0.1619	(0.0025)	**0.8654***	(0.0091)	0.2727	(0.0039)
ECL1	0.5102	(0.0020)	0.4747	(0.0012)	0.1527	(0.0014)	0.8165	(0.0122)	0.2573	(0.0025)
JR4	0.7937	(0.0349)	0.8383	(0.0433)	0.2192	(0.0228)	0.3939	(0.0551)	0.2789	(0.0192)
JR5	0.8070	(0.0120)	0.8423	(0.0093)	0.2556	(0.0516)	0.4915	(0.1505)	0.3356	(0.0796)
JR6	0.7231	(0.0021)	0.7175	(0.0010)	0.2406	(0.0021)	0.7717	(0.0123)	0.3667	(0.0037)
PH2	0.8813	(0.0152)	0.9120	(0.0164)	0.4446	(0.0396)	0.6104	(0.0095)	**0.5134**	(0.0275)
SSKS1	**0.9007**	(0.0027)	**0.9383**	(0.0009)	**0.4601**	(0.0099)	0.5259	(0.0172)	0.4908	(0.0131)
TCCP1	0.8563	(0.0006)	0.9147	(0.0005)	0.2560	(0.0043)	0.2862	(0.0039)	0.2703	(0.0041)
TCCP2	0.8561	(0.0007)	0.9150	(0.0005)	0.2587	(0.0053)	0.2867	(0.0039)	0.2720	(0.0046)
SBC1	0.8483	(0.0025)	0.8708	(0.0042)	0.3703	(0.0062)	0.6550**	(0.0141)	0.4730	(0.0044)
SC1	0.8501	(0.0018)	0.8737	(0.0033)	0.3723	(0.0051)	0.6460**	(0.0099)	0.4723	(0.0015)

Table 5: Comparative results (means and standard deviations) for Binary Relevance, using the Mirex'09 Mood Tag Dataset. The best results are indicated in bold. *The high recall in this case is artificial, given the low precision achieved. ** If we take into account both precision and recall, we can see that our approach achieves a good recall without hampering somehow the precision

F-measure by Tag



Significance tests

* **F-measure by Fold**: No statistical significance between all the presented algorithms.

* **F-measure by Tag**: (PH2, SSKS1, BAx)>(SC1, SBC1) are statistically significant in MajorMiner only, no statistical significance between them in the Mood Tag dataset. SC1, SBC1>TCCPx statistically significant in MajorMiner. SBC1>SC1 not statistically significant.

Conclusions

* Our algorithm has shown to be:
  - *Fast*: up to 2-3 minutes to train and classify a fold.
  - *Scalable*: in SC1, for example, each audio excerpt is defined by a single vector of 29-30 dimensions.
  - *Easy to implement*: built on top of a k-NN classifier, defines simple heuristics for classification and affinity.
  - *Consistent*: the overall results, averaged per fold, show a very low standard deviation (see, for instance, how PH2 performs in different folds). Furthermore, the algorithm ranks between second and fourth (out of 15 participants) in almost all the evaluation results.

* Statistical significance was found mainly in MajorMiner at tag level, probably due to the diverse (and correlated) concepts/categories used in the Tag vocabulary. Yet, in many evaluation results there was no statistical significance. This issue is not new in MIREX Audio tag classification.

* After 3 years with the same set of evaluation measures, are all of them good and sufficient to discriminate between different algorithms?

* Are we actually reaching a "glass ceiling" for audio tag classification, even with novel algorithms that take into account correlations between higher level concepts?

References

[1] M. Sordo, C. Laurier, and O. Celma "Annotating Music Collections: How content-based similarity helps to propagate labels". Proceedings of the International Conference on Music Information Retrieval, pp. 531–534, 2007.

[2] D. Bogdanov, J. Serrà , N. Wack, P. Herrera and X. Serra “Unifying Low-level and High-level Music Similarity Measures", IEEE Transactions on Multimedia, vol. 13, pp. 587–701, 2011.