Active Mining of Cell Assay Images

An active learning approach with cluster analysis

Nicolas Cebron

GK Spring Workshop

March 16, 2005

- Cohn, David: Improving Generalization with Active Learning, 1992
- Warmuth et. al.: Active Learning with Support Vector Machines in the Drug Discovery Process, 2002

What is Active Learning ?

- Large dataset with unlabeled data
- Normally, training examples are choosen at random.
- Active Learning: algorithm has control over the input it trains on.
- Learner queries a point in the input domain $\rightarrow$ teacher/oracle returns classification of this point.

Region of uncertainty

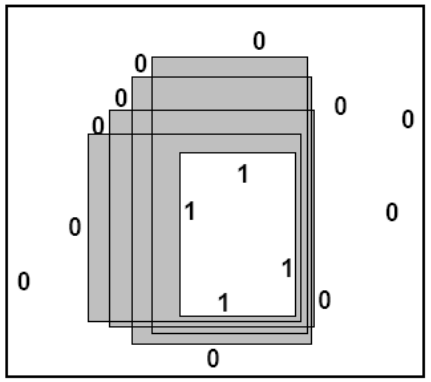


Figure: $R(S^m)$ shaded

- Concept c : subset of points in the domain (e.g. rectangle)
- Set S^m of m examples
- Areas not determined by available information: $\mathrm{R}(\mathrm{S}^m) = \{x : \exists c_1, c_2 \in C, c_1, c_2$ are consistent with all $\mathrm{s} \in S^m$ and $c_1(x) \neq c_2(x)$.

Region of uncertainty (cont.)

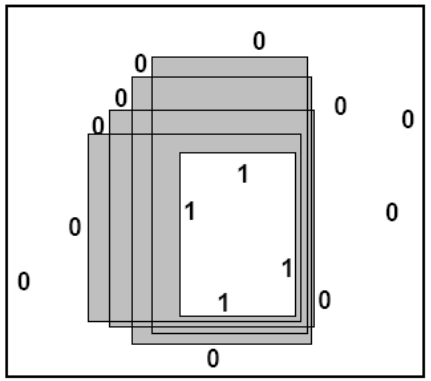


Figure: $R(S^m)$ shaded

- Draw at random over the whole domain
- Most examples we draw (outside $R(S^m)$) will not provide us with information about the concept we are trying to learn
- Points outside $R(S^m)$ leave $R(S^m)$ unchanged, points inside will further restrict the region.
- Any disagreement between concepts must lie within $R(S^m)$.

Selective sampling

- Draw samples only from within $R(S^m)$.
- Draw an unclassified example, query the classification.
- Recalculate $R(S^m)$ after each new example.
- Recalculating $R(S^m)$ may be computationally expensive $\rightarrow$ perform selective sampling in batches.

Introduction

Goal of cluster analysis:

- Group a set of objects into homogenous groups.
- find dense and sparse regions in the dataset.
- find patterns in the underlying data.

Fuzzy c-means

Given:

- Data record $\{x_1, \ldots, x_n\} \in \Re$
- Fixed number of clusters (prototypes) c

Objective function:

$$f = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} d_{ij}$$

- u_{ij} = degree of membership of record x_j to cluster i
  $u_{ij} \in [0, 1]$ and $\sum_{i=1}^{c} = 1, \forall j = 1, \ldots, n$
- d_{ij} = distance of object x_j to cluster i
- Fuzzifier m indicates how much clusters can overlap

Fuzzy c-means (cont.)

Non-linear optimisation problem, therefore partial optimisation:

- Cluster as centroid of his member-datarows:

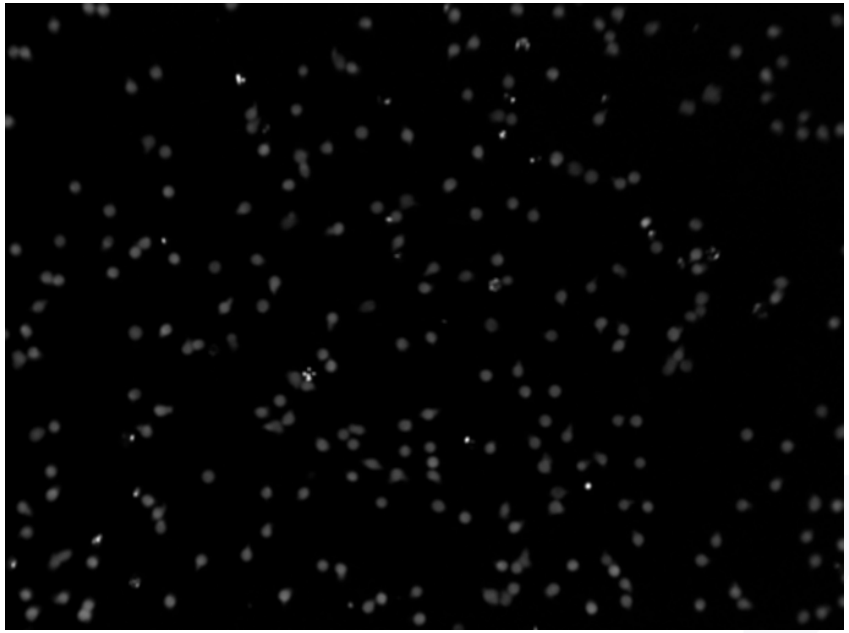
$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}$$

- New memberships:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{\frac{2}{m-1}}}$$

Cell Assay Images

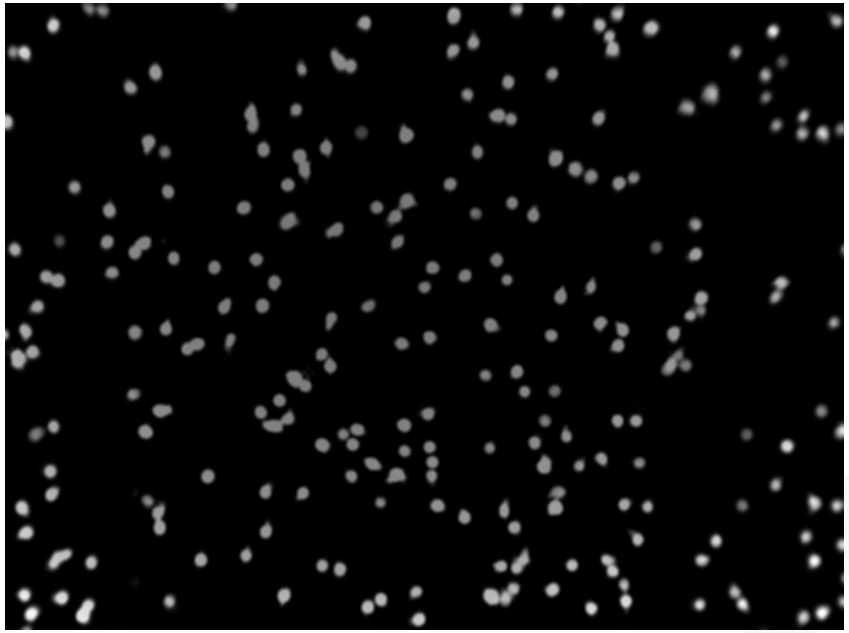
- Images obtained by a fluorescence microscope camera.
- Cells are treated with an agent.



Before

Cell Assay Images

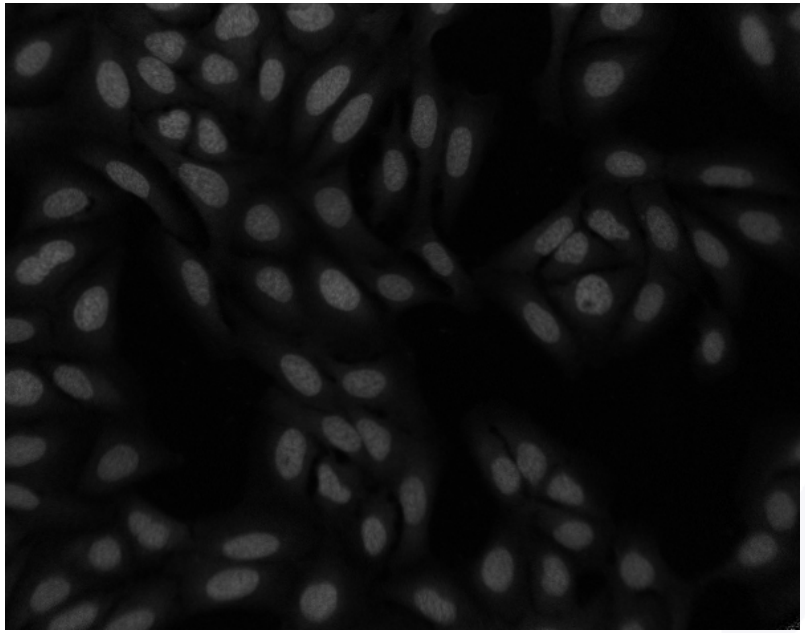
- Images obtained by a fluorescence microscope camera.
- Cells are treated with an agent.



After

Cell Assay Images

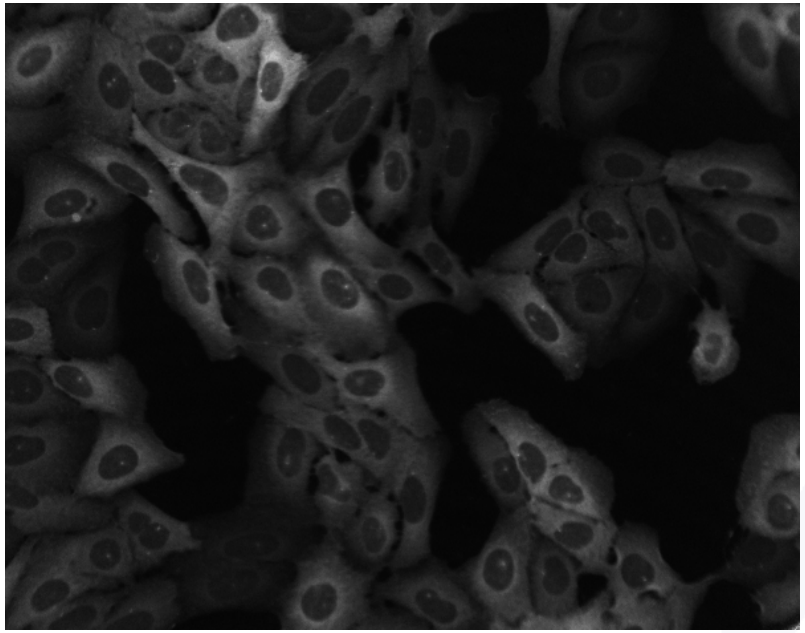
- Images obtained by a fluorescence microscope camera.
- Cells are treated with an agent.



Before

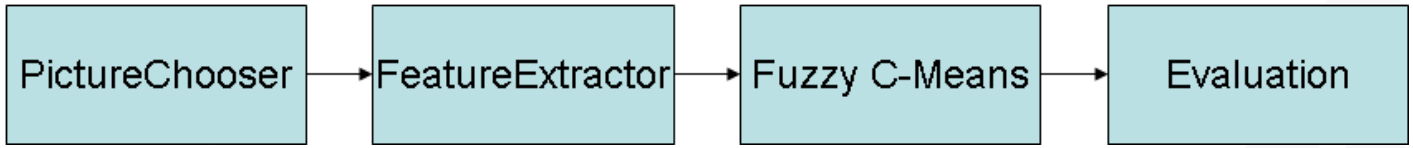
Cell Assay Images

- Images obtained by a fluorescence microscope camera.
- Cells are treated with an agent.

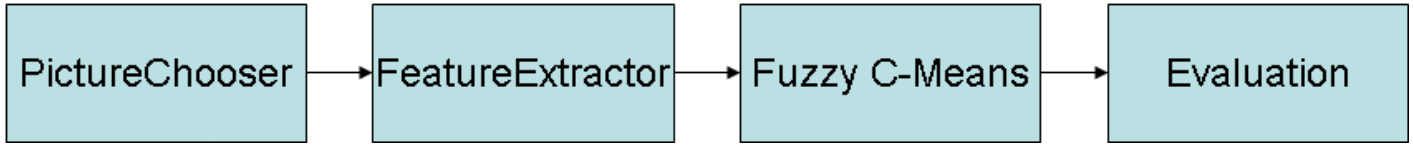


After

Workflow


PictureChooser
FeatureExtractor
Fuzzy C-Means
Evaluation

Workflow - PictureChooser



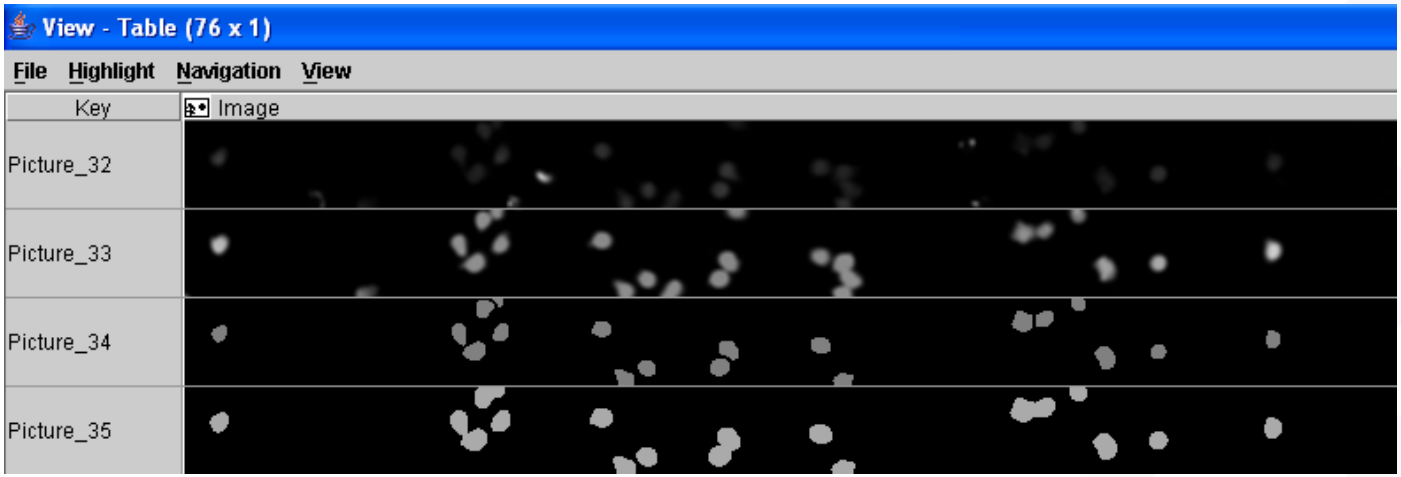
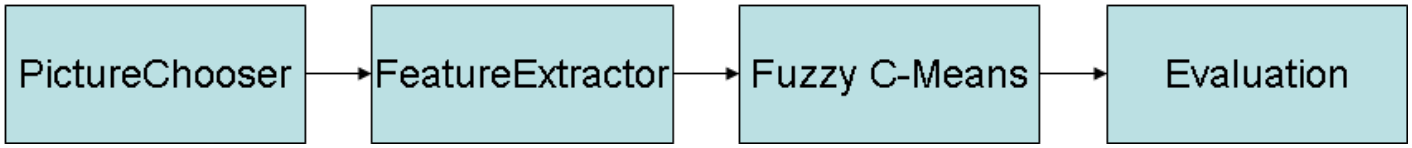


Figure: Pictures obtained from PictureChooser

Workflow - FeatureExtraction



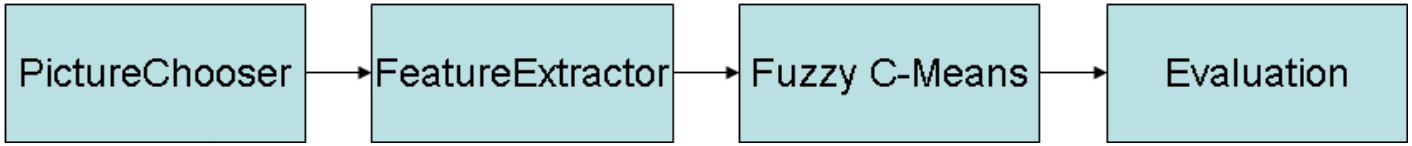
View - Table (101+ x 16)

File Highlight Navigation View

Key	Identifier	Before	Mean	StdDev	Mode	Max	After	Mean_a	StdDev_a	Mode_a	Max_a	Diff_Me...	Diff_Std...	Diff_Mo...	Diff_Max	Preclas...
11012001_0...	11012001_...		295.838	128.885	31	511		4673.25	1841.719	124	7471	4377.412	1712.834	93	6960	2
11012001_0...	11012001_...		202.475	84.079	244	336		3118.083	1298.606	130	5208	2915.608	1214.527	-114	4872	2
11012001_0...	11012001_...		294.933	131.676	139	497		3844	1357.207	250	5394	3549.067	1225.531	111	4897	2
11012001_0...	11012001_...		269.412	97.804	161	420		3875.388	1278.196	140	5549	3605.975	1180.392	-21	5129	2
11012001_0...	11012001_...		185.811	101.093	75	364		3832.289	1522.508	125	6045	3646.478	1421.416	50	5681	2
11012001_0...	11012001_...		237.222	109.178	22	434		3966.747	1369.875	248	5394	3729.525	1260.697	226	4960	2
11012001_0...	11012001_...		198.475	83.578	62	329		4547.606	1519.552	254	6417	4349.131	1435.973	192	6088	2
11012001_0...	11012001_...		226.227	103.022	86	413		3616.573	1433.9	252	5332	3390.345	1330.877	166	4919	2
11012001_0...	11012001_...		204.488	74.654	141	315		4529.875	1488.265	253	6355	4325.388	1413.611	112	6040	2
11012001_0...	11012001_...		144.285	85.821	29	301		3473.691	1566.842	249	5146	3329.406	1481.021	220	4845	2
11012001_0...	11012001_...		175.622	74.108	134	294		4025.867	1545.627	105	6355	3850.244	1471.519	-29	6061	2
11012001_0...	11012001_...		166.091	73.864	134	294		2886.758	1328.145	211	5208	2720.667	1254.282	77	4914	2
11012001_0...	11012001_...		102.667	51.849	105	210		2439.442	1222.242	172	4588	2336.775	1170.393	67	4378	2
11012001_0...	11012001_...		151.693	58.506	90	245		2706.511	1105.812	87	4650	2554.818	1047.306	-3	4405	2
11012001_0...	11012001_...		179.978	76.2	243	287		3014.922	1378.767	228	4960	2834.944	1302.567	-15	4673	2
11012001_0...	11012001_...		74.136	46.585	0	154		3614.788	1717.916	95	6665	3540.652	1671.331	95	6511	2

Figure: Features based on histogram

Workflow - FeatureExtraction



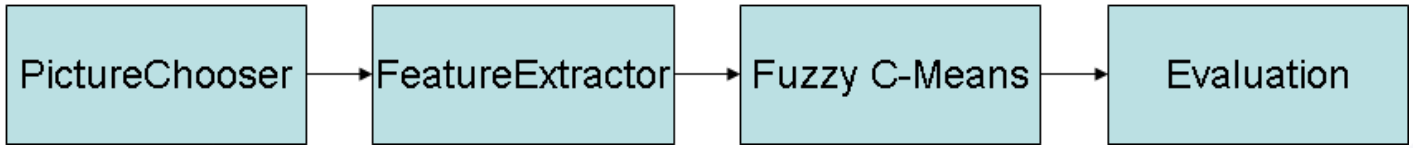
View - Table (101+ x 14)

File Highlight Navigation View

Key	Identifier	Before	bp[0]	bp[1]	bp[2]	bp[3]	bp[4]	After	ap[0]	ap[1]	ap[2]	ap[3]	ap[4]	Preclas...
19006001_0...	19006001_...		9	108	432	486	324		70	2695	5285	5250	4445	2
19006001_0...	19006001_...		18	63	171	216	225		840	2695	5705	5845	4025	2
19006001_0...	19006001_...		9	117	306	333	198		350	2065	4200	4235	2030	2
19006001_0...	19006001_...		45	513	684	522	180		525	3605	4375	2520	805	2
19006001_0...	19006001_...		171	261	378	504	495		4025	4970	5250	5250	5215	2
19006001_0...	19006001_...		126	207	378	540	540		3045	4095	5810	6650	6650	2
19006001_0...	19006001_...		18	144	441	576	558		840	3115	5460	5670	5390	2
19006001_0...	19006001_...		90	297	468	459	225		350	4900	5040	5005	2555	2
19006001_0...	19006001_...		90	315	585	675	549		2520	4970	5460	5390	5215	2
19006001_0...	19006001_...		18	153	360	414	342		105	2170	4935	4935	4830	2
19006001_0...	19006001_...		54	270	441	531	432		2380	5740	7140	7245	5180	2
19006001_0...	19006001_...		18	126	378	414	180		840	3990	6055	5985	2380	2
19006001_0...	19006001_...		297	333	450	423	225		1960	5285	5460	5495	2940	2
19006001_0...	19006001_...		27	117	297	369	324		245	2870	5355	5320	5180	2
19006001_0...	19006001_...		72	108	324	378	351		1645	1960	4270	4620	3395	2
19006001_0...	19006001_...		0	108	369	558	351		175	2590	5845	6055	3115	2
19006001_0...	19006001_...		0	117	342	378	297		210	3255	4830	4865	4690	2
19006001_0...	19006001_...		45	225	423	423	207		490	4480	5530	5495	3500	2
19006001_0...	19006001_...		63	162	234	306	297		1050	3850	5810	7350	7140	2
19006001_0...	19006001_...		234	261	414	441	369		3605	5110	6825	6685	5145	2
19006001_0...	19006001_...		45	243	414	468	351		665	3115	4655	4795	3500	2

Figure: Features based on feature vector

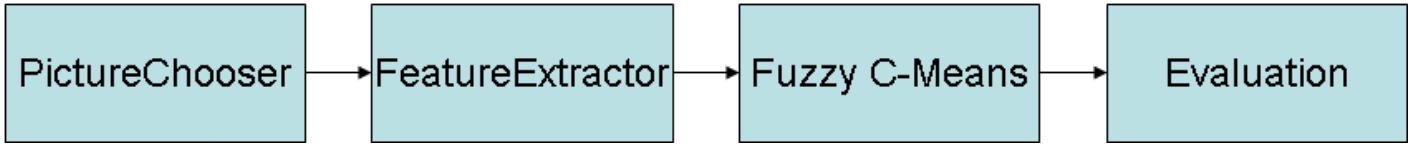
Workflow - Fuzzy c-means



Key	Identifier	Before	bp[0]	bp[1]	bp[2]	bp[3]	bp[4]	After	ap[0]	ap[1]	ap[2]	ap[3]	ap[4]	Preclas...	Cluster0	Cluster1	Cluster2	Cluster3	Cluster4
122.0	17013001_...		72	324	432	531	432		936	4251	4680	4641	4446	2	71%	17%	8%	3%	1%
2131.0	20014001_...		63	273	455	511	483		684	3492	4680	4680	4716	2	71%	21%	5%	2%	1%
2123.0	20014001_...		126	385	511	518	462		1116	4140	4824	4788	4752	2	72%	14%	10%	3%	1%
3659.0	17012001_...		49	238	399	448	385		578	3570	4964	4964	4998	2	72%	16%	10%	2%	1%
2106.0	20014001_...		35	238	371	399	427		864	3924	5148	5148	4968	2	72%	11%	14%	2%	1%
3235.0	14004001_...		72	288	414	432	396		783	3596	4582	4611	4611	2	72%	20%	5%	2%	1%
2816.0	21013001_...		169	312	416	429	273		1564	3910	5134	5168	4556	2	72%	12%	13%	2%	1%
1486.0	17003001_...		132	450	636	582	264		1120	3968	4832	4832	3840	2	72%	18%	6%	3%	1%
629.0	13007001_...		70	270	390	420	380		928	3200	4704	4800	4320	2	72%	23%	3%	2%	0%
3616.0	17012001_...		91	231	343	343	315		986	3196	5100	5134	4896	2	72%	16%	9%	2%	1%
2027.0	20014001_...		56	287	420	469	448		540	3528	5004	4968	4932	2	73%	16%	9%	2%	1%
113.0	17013001_...		63	153	306	378	342		507	3744	4680	4641	4485	2	73%	19%	5%	2%	1%
559.0	13007001_...		40	160	310	310	310		1376	3616	5184	5152	4832	2	73%	12%	12%	2%	1%
3244.0	14004001_...		90	234	306	324	198		464	3799	4930	5046	3799	2	73%	18%	5%	3%	1%
104.0	17013001_...		63	207	360	378	333		1365	3627	4641	4680	4563	2	73%	18%	6%	3%	1%
3567.0	17012001_...		42	203	322	329	301		612	3264	4964	4998	4828	2	73%	18%	6%	2%	0%
216.0	17013001_...		18	180	324	342	234		585	3978	5304	5421	3939	2	73%	14%	10%	3%	1%
2373.0	11013001_...		0	203	348	290	232		884	4284	4760	4794	4522	2	73%	14%	9%	3%	1%
633.0	13007001_...		80	250	340	380	280		1248	3808	5056	5216	3680	2	74%	16%	7%	3%	1%
3669.0	17012001_...		7	140	343	364	238		340	3502	4998	5032	4692	2	74%	16%	7%	2%	1%
571.0	13007001_...		70	220	340	350	330		736	3648	4640	4608	4608	2	74%	18%	5%	2%	1%
2662.0	21013001_...		39	169	260	312	260		1326	3774	5066	5508	4590	2	74%	11%	12%	2%	1%
1617.0	17003001_...		84	312	522	600	552		800	3552	5408	5440	4064	2	75%	13%	9%	2%	1%
2655.0	21013001_...		39	286	442	377	143		510	4148	4896	4828	4250	2	75%	15%	7%	3%	1%
779.0	7014001_0...		28	203	322	350	329		884	4046	4726	4692	4692	2	75%	14%	7%	3%	1%
2369.0	11013001_...		0	87	174	261	232		986	3536	4726	4760	4760	2	75%	16%	6%	2%	1%

Figure: Cluster0 attracts positive cells

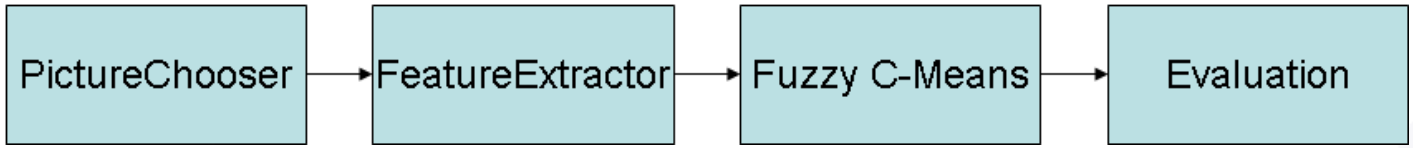
Workflow - Fuzzy c-means



Key	S Identifier	Before	D bp[0]	D bp[1]	D bp[2]	D bp[3]	D bp[4]	After	D ap[0]	D ap[1]	D ap[2]	D ap[3]	D ap[4]	I Preclas...	D Cluster0	D Cluster1	D Cluster2	D Cluster3	D Cluster4
362.0	6007001_0...		13	143	312	390	403		60	420	900	1140	1140	1	1%	2%	1%	6%	90%
1939.0	3003001_0...		72	486	756	828	774		48	448	848	960	928	1	1%	2%	1%	6%	90%
2507.0	5012001_0...		96	312	606	624	528		234	648	1116	1098	828	1	1%	2%	1%	6%	90%
1788.0	3003001_0...		522	198	252	198	18		624	192	320	256	16	1	2%	2%	1%	5%	90%
1646.0	5011001_0...		42	168	392	420	378		68	374	1088	1190	1156	1	2%	2%	1%	8%	87%
451.0	6007001_0...		78	481	858	975	949		100	520	860	940	840	1	2%	3%	1%	7%	87%
2494.0	5012001_0...		30	270	456	552	528		54	630	1044	1188	1170	1	2%	3%	1%	9%	85%
1622.0	5011001_0...		518	406	490	574	560		935	765	442	629	629	1	2%	3%	1%	7%	87%
1820.0	3003001_0...		18	306	648	990	954		80	528	912	1120	1040	1	2%	3%	1%	9%	85%
2585.0	5012001_0...		180	348	600	636	510		126	468	1242	1278	1170	1	2%	3%	1%	11%	81%
2497.0	5012001_0...		48	252	348	372	396		72	702	1152	1332	1350	1	3%	4%	2%	15%	77%
1663.0	5011001_0...		70	224	462	490	462		85	459	1360	1479	1360	1	3%	5%	2%	18%	72%
1681.0	5011001_0...		28	294	476	518	490		68	799	1428	1445	1326	1	4%	5%	2%	21%	68%
1779.0	5011001_0...		42	322	560	560	518		119	952	1530	1615	1479	1	5%	7%	3%	29%	56%
3347.0	14004001_...		162	504	756	1152	1800		174	696	1044	1305	1015	1	5%	7%	3%	18%	67%
2483.0	5012001_0...		72	240	462	498	456		198	738	1638	1800	1602	1	5%	8%	3%	36%	49%
300.0	6007001_0...		52	1898	3315	3315	312		80	3100	4680	4260	80	1	22%	24%	14%	25%	15%
1925.0	3003001_0...		180	1674	4590	4590	4482		336	1936	4080	4080	3744	1	22%	24%	18%	21%	15%
276.0	6007001_0...		169	884	1326	1391	962		600	2880	4340	4240	2800	1	27%	40%	9%	20%	4%
521.0	13007001_...		190	220	230	190	170		3008	4384	5600	4320	3456	1	40%	25%	21%	11%	4%
1012.0	11010001_...		36	99	162	162	117		528	1656	2928	3216	2520	2	0%	1%	0%	99%	0%
3176.0	17010001_...		16	68	160	180	160		396	1584	2904	3146	2728	2	0%	1%	0%	99%	0%
2594.0	5012001_0...		6	30	144	216	216		18	252	522	558	522	2	0%	0%	0%	1%	98%
1749.0	5011001_0...		0	56	224	406	462		170	459	680	850	714	2	0%	1%	0%	1%	97%
1651.0	5011001_0...		0	168	280	210	126		0	34	765	867	340	2	0%	1%	0%	2%	97%
1074.0	11010001_...		36	108	207	216	162		408	1392	2952	3024	2376	2	0%	1%	0%	98%	0%
931.0	11010001_...		27	72	153	144	99		480	1512	3096	3144	2616	2	1%	1%	0%	98%	0%
1639.0	5011001_0...		0	28	238	224	196		0	34	629	850	782	2	1%	1%	0%	2%	96%
2867.0	9009001_0...		35	161	259	280	210		360	1760	3000	3260	2620	2	1%	1%	0%	97%	0%
1042.0	11010001_...		9	90	144	144	126		192	1704	3024	3000	2616	2	1%	1%	0%	97%	1%
1065.0	11010001_...		27	126	189	207	207		240	1680	2856	3024	2760	2	1%	1%	0%	97%	1%
1056.0	11010001_...		9	81	126	162	162		264	1776	2664	3168	2472	2	1%	1%	0%	97%	1%

Figure: Outliers and false preclassifications

Workflow - Evaluation



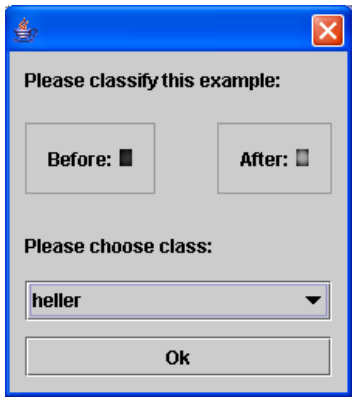


Figure: Expert Interaction

Results

Pre-classification	sure		unsure	
	darker (1)	brighter (2)	darker (1)	brighter (2)
1	17.8 %	0.02 %	0 %	0 %
2	4.01 %	74.9 %	1.4 %	1.71 %

Future Directions

Feature extraction (more sophisticated):

- DFT (Texture features)
- Shape
- Based on existing library of subroutines
- Selection of the most useful features (use subspace of original feature space)

Active Learning:

- Fit the clusters (not only to match the underlying data structure but also to match the classification task)
- Integrate Expert Risk Assessment (e.g. number of confidently classifed images vs. additional labelling work)

The End

Thank you for your attention !