# A Label Ranking extension for Weka

## Alexander Balz

February 26, 2013

#### 1 Introduction

This document describes a modification for the machine learning framework Weka[1], in which Weka will be able to process Label Ranking data sets and to predict rankings by using the currently implemented "*Ranking by Pairwise Comparison*" (RPC) algorithm[2] or the Decision Tree based method *TreeLabelRanking* [3].

In the following, the .xarff file type will be introduced, which simply is an extension of the existing .arff file type supported by Weka. Next, the differences to the original Weka preprocess panel are discussed. Afterwards, it is shown how label rankings can be predicted from a given data set and which statistics are returned.

## 2 File format of Label Ranking data sets

Label ranking data sets can be saved in .xarff file format. The name already implies that it is an extension of Weka's regular .arff files[1]. By giving an .xarff file fragment, we will explain the differences of this file type to .arff files:

```
1
       @relation example
\mathbf{2}
       @attribute A1 NUMERIC
3
       @attribute A2 NUMERIC
4
       @attribute A3 NUMERIC
5
       @attribute A4 RANKING {L1,L2,L3,L4,L5}
6
7
       @data
8
       -1.337785, 1.038478, 1.856137, L5>L4>L3>L2>L1
9
       -1.237785, 1.033796, 1.956137, L4>L2>L3
       -1.327803,\, 1.052523,\, 0.982119,\, {\tt L5>L4>L1|L3>L1|L2>L4}
10
       -0.124868, 9.354512, 1.112111, 'L1 > L2 > L3'
11
```

In line 5, a ranking attribute has been defined. Its structure is quite similar to nominal

attributes. A ranking attribute needs to have a unique attribute name, followed by the keyword 'RANKING'. After this key, a list of label names is following. Label rankings must only consist of those labels included into the list. For each instance, pairwise comparisons, partial or complete rankings can be assigned. In more detail: Labels are assembled to a Label Ranking by using the '>' character. So Label1>Label2 means that label 1 is preferred over label 2. It is not permitted to state one single label inside of an instance, so there have to be at least two labels separated by the '>' sign. If you wish to use whitespaces inside of the rankings, please make sure that they must be written with apostrophes surrounding, like 'L1 > L2 > L3'. In case of non-existant whitespaces, there is no need to put rankings between apostrophes.

Some examples are given from line 8 on. First, a complete ranking is shown - consisting of all declared labels. Nevertheless, label rankings do not necessarily have to include all labels available in the ranking attribute, so that such rankings like in line 9 are permitted, too. Furthermore, it is possible to insert partial label rankings (line 10). In order to do so, several rankings are separated by the '|' symbol. In line 11, there is an example for a ranking containing whitespaces.

## 3 Processing .xarff files

.xarff files can be loaded - like the .arff files inside of the preprocess panel of Weka's explorer. After having selected the right file extension and loading a desired file, Weka will display attribute statistics as usual. When clicking on the ranking attribute, one can see a matrix inside of the 'Selected attribute' panel on the right, containing the amount of every pairwise label lanking that occurs inside of the data set. Under this matrix, a visualization of it is displayed. A white field means that no or hardly any pairwise ranking is existant into the loaded file. The darker a matrix field appears, the more pairwise rankings are contained in the data set.

So the preprocess panel could look like in figure 1.

ed attribut me: L RAN ing: D (DN) L1 L2	Unde Kans	Dat	[	Edt			Save	Acol	
ed attribut me: LRAN ing: 0 (DN) L1 L2	Unde E KONG	Dat		Edt			Save	Acci	
ed attribut ne: LRAN ng: D (D% L1 L3	e KING J	Dat						Acel	
ed attribut me: LRAN ing: D (D% L1 L2	ini KING V	Dat						Acel	1
ne: LRAN ne: LRAN ng: D (D%	e KINS J	Dat							8
ne: LRAN ing: D (D%	KING V	Date							
L1 L2			nct: 0			Type: R Nove: 0	anking (0%)		
0 151	2 1.3	14	1.5	1.6	17	LB	1.9	L10	
N 33	88 8740	8508	8825	1428	9023	5980	1752	4042	١.
5804 0	8513	9097	9136	4146	6610	6320	5193	3949	ĥ
2252 24	(79 0	\$000	5061	1456	4686	2425	1511	2159	1
2404 10	95 5992	0	4879	1548	4013	3041	1831	1175	]1
2367 18	155 5931	6013	0	1239	3792	2391	1389	1378	1
9964 69	146 9536	9-944	9753	0	9958	9432	7964	6892	Н
1969 43	82 6306	6979	7200	2034	0	2857	1065	2968	Į.
5012 46	72 8567	7951	8901	2560	8135	0	3514	3886	1.
RANKING	(Rnk)						* N	science	Л
							CO.		-
L1 L2	1.3	14	1.5	16	17	1.8	1.9	L10	
								1	
								1	
	1			2				-	
			_						
_									
_					_			_	
				-		-		-	
	and the second second						-		
			_			-	-		
D									

Figure 1: The Preprocess Panel including Label Ranking data

# 4 Editing data sets manually

In Wekas preprocess panel, it is possible to change entries of the read data set by clicking on the "Edit" button. In order to do so, double click on the value that should be changed and type in the desired value or ranking.

Complete and incomplete rankings may be written into the fields for ranking attributes, see figure 2.

s. V	liewer							25				= = ×
Relati	on: cpu-sr	nal_dense										
No.	1: A1 Numeric	2: AZ Numeric	3: A3 Numeric	4: A4 Numeric	St AS Numeric	6: A6 Numeric	7: L RANKE Ranking	NG	b			
1	-0.254	-0.371	-0.777	-0.540	-0.224	-0.356	1182833284358			Edit	ا لحد	Save
2	-0.347	-0.438	-0.086	-0.02754	-0.305	-0.598	11×12	-				
3	0.795	2.137	0.918	0.238	0.099	-0.195	145.1512513515					
4	-0.27299	-0.438	1.47812	0.228	0.167	-0.360	13×1>12×4×15					Apply
5	0.420	1.401	1.005	0.193	0.585	0.288	L1>L2 L4>L5 L2>L3					
6	-0.27289	-0.405	-0.376	-0.394	0.392	-0.598	15×12×13×14×11					
7	-0.310	-0.438	·1.023	-0.731	0.642	0.449	11>14>13>12>15				Type	: Ranking
8	-0.235	-0.271	-0.590	0.148	0.187	-0.598	15×12×1>14×13		Distri	ct: 0	Unique	: 0 (0%)
9	2.613	0.899	0.083	0.444	0.697	-0.356	14×12×1×13×15		2	13	1.4	15
10	-0.347	-0.438	0.621	0.304	0.124	-0.356	13>13>13>13>13>13>14>15		1.224	19949	194.99	4510
11	-0.197	-0.338	0.416	-0.525	0.31193	-0.276	14x1x12x13x15		001	3393	30/0	4219
12	-0.160	-0.371	-0.379	-0.675	0.21846	-0.598	14×15×11×12×13		4.79	0	3/76	7/21
13	-0.329	-0.405	1.894	0.982	0.772	-0.679	1301501012014		100	2720	4466	3304
14	-0.254	-0.371	-0.163	-0.449	0.237	-0.518	14×15×13×1×12		100	3//0	2610	4003
15	2.857	3.341	-0.538	-0.198	0.100	-0.518	13>12>14>11>15		24/1	1000	3239	0
16	1.170	1.033	0.021	-0.273	1.07766	-0.518	14>13>12>11>15					
17	0.32688	1.200	1.445	2.324	1.91266	0.288	14>12>11>13>15		-			
18	-0.122	-0.037	0.338	1.248	0.404	-0.518	14912911913915					<ul> <li>Visualize Al</li> </ul>
19	0.083	0.632	-1.1553	-0.851	0.765	0.04656	13015014012011					1.5
20	0.889	2.572	-0.27933	-0.399	-0.554	-0.276	11×5×4×2×3			L3	L4	15
21	-0.085	-0.271	-0.312	0.665	0.367	-0.356	13>14>11>12>15				_	
22	-0.291	-0.438	-0.206	-0.007	0.311	0.207	15×13×1×1×12					Contraction of the
23	-0.179	-0.30464	0.873	0.007	0.111	-0.276	15×12×1>14×13					A CONTRACTOR OF THE OWNER.
10	1	1			10							
												The second s
							Undo OK	Cancel				Concession of the local division of the loca
L				Remove			L5					
Stat	tus										Log	x

Figure 2: The editing viewer

# **5** Predicting Label Rankings

After having loaded the data set, you are able to make a prediction over it. A ranking algorithm can be chosen in the "Classify/Rank" panel of the Weka explorer. Within the dialog for choosing a Classifier, we can find a folder named labelranking. Inside of it, the ranking by pairwise comparison (RPC) algorithm can be found [2]. This algorithm has several parameters which can be changed by clicking on its name in the Classifier/Label ranker field (figure 3).

eka.dass	ifiers.labelranking.RPC
About	
Label F	Ranking by pairwise comparison. More
	Capabilities
classifier	Choose Logistic -R 1.0E-8 -M -1
classifier debug	Choose Logistic -R 1.0E-8 -M -1 False
classifier	Choose Logistic -R 1.0E-8 -M -1

Figure 3: The parameters of RPC

RPC needs a base classifier to work properly. The default base classifier is logistic regression. Next, the debug parameter is contained in most classifiers inside of Weka and provides additional information on the console. RPC does a voting on the base classifiers' results. There are two different ways of voting: The first way of voting is *binary voting*, in which the base classifiers return their predictions for every pairwise label ranking. The values of those predictions may be 0 or 1. Aftwerwards, the labels with the highest amount of votes are proposed as final prediction. The second voting scheme is called *soft voting*. In this voting method, the base classifiers return a real number between 0 and 1 and including for each pairwise ranking. The closer this value to 1, the stronger it is taken into consideration in the voting process. For more details, see [2].

The default value of this parameter is soft voting. The results after a run of the RPC algorithm look similar to figure 4.

Some ranking measures have been implemented to evaluate the quality of prediction. These are Spearman footrule, Spearman rank correlation and Kendall's tau. Please note that those measures will only be shown if all rankings inside of the data set are complete rankings over a subset of the data set's available labels. Thus, if one or several partial rankings appear inside of the data set, the aforementioned measures will not be calculated.

Additionally, a matrix is computed which shows us the relative error of all pairwise label rankings. All pairwise rankings have been derived of the data set and are compared to those of the predictions. Correct and incorrect predictions are summarized into the matrix, separated by the '|' sign. Below, the summed up correct/incorrect values are returned. Unlike the ranking measures, this matrix will be shown even if partial rankings occur inside of the instances.

It is not recommended to use the functions of cluster and associate panels on label ranking data, because no algorithms are provided for this use yet and the existing ones

Country (and value)         Country (and value)           Channel With one of the Markows Laples = 4 1000 4 M - 3           Statigation         Country (and value adjust)           Country (and value adjust)	Preprocess County, Runk Queter Associa	Az Sele	ct attributes	Hausida		_				-
Observe         Standard out or supplet           Out Standard out or supplet         Construction of the supplet           Out Standard out or supplet         Standard out or supplet           Out Standard out or supplet         Standard out or supplet           Out Standard out or supplet         Standard out or supplet           Out Standard out or supplet         Standard out or supplet           Out Standard out or supplet         Standard out or supplet           Out Standard out or supplet         Standard out or supplet           Out Standard out or supplet         Standard out or supplet	Casa/lan Label varias									-
Status         Construction ratio         Construction ratio           Out-based function         Social function ratio         Social function         Social function           December 1         Social function         Social function         Social function         Social function           Presenting out         Social function         Social function         Social function         Social function           Presenting out         Social function         Social function         Social function         Social function           Presenting out         Social function         Social function         Social function         Social function           Presenting out         Social function         Social function         Social function         Social function           Presenting out         Social function         Social function         Social function         Social function           Presenting out         Social function         Social function         Social function         Social function           Presenting out         Social function         Social function         Social function         Social function           Presenting out         Social function         Social function         Social function         Social function           Social function         Social function         Social function         Social	Chane RPC -918 -0 8 -W webs class	hes.hav	tions Logette	- #.1.00-0 4	N-1					3
Over Standard Instant, Son, Son, Son, Son, Son, Son, Son, Son	Text options	Casafe	-Jakel ranke	Inglue						
Series         Series<	O the transport Saskel last set Gross-widden Peids Proventage set Hore sptane	ann 3 ann 3 Speac Speac Randa	Annual field man foote man correl d1's tau	dross-vel arunes ule lation	16.227 0.742 0.648					
Sert         Ymp         3         12         13         14         15         14         15           Read bit (parted & region)         11         101         1241129         22642         1251127         145146         1264161         15614           12         1411         1441129         22642         1251127         145146         156141         15614           12         1411         1491         1491         1491         156141         15614           12         1411         2114         12112         1614         156145         156145           12         1411         2114         121112         1614         156145         156145           12         1411         21144         158144         1191415         168145         156145           12         12110         1411124         119141         119145         168145         168145           12         12110         1411124         119141         168145         168144         119144         168144           12         12110         141142         158145         168144         168144         168144         168144         168144         168144         168144         168144	Bolis +		METRONIAN C	neer of pe	redictions					
incorrect: 1705	Sant Ino Realt At Lyter dat for spine() 1435555 Met wing SFC	> 11 12 13 14 15 14 13 14 15 14	L1 010 L45129 41027 52178 13129	12 134(22) 414 251242 52(57) 134(24) 52(57) 134(14) 135(150) 134(14) 135(150) 134(14) 136(15) 134(14) 136(15)1	13 220142 200111 010 11211210 1521104 201104 1551214 1651214 165124 165124 165124 165124 165124 165124	L4 155:127 220:04 150:126 0:0 120:128 140:104 153:120 0:120 0:120 0:120 0:120 0:120 0:120 0:120 0:120 0:120 0:120 0:120 0:120 0:120 0:120 0:127 0:127 0:0 120:127 120:04 120:127 120:04 120:127 120:04 120:127 120:04 120:127 120:04 120:127 120:04 120:127 120:04 120:127 120:04 120:127 120:127 120:04 120:126 120:1	15 145144 174189 187145 198179 610 129104 1091200 821209 42199 25105	54 204049 154041 152140 1640240 57076 24040 24040 24040 22040	17 152123 158120 158150 158150 158150 158150 248150 810 11140 24935 11140 24935	
(x) 0		×					1			

Figure 4: The Classifier Panel after RPC has been executed

inside of these panels will mostly cause errors when trying to evaluate them on Label Ranking data sets.

Using the tree label ranking (TLR) algorithm works quite similar to RPC, but TLR is no meta-classifier, so choosing a base classifier is not necessary.

The output has been slightly extended. A summary of the generated tree will be shown after the classifying process in figure 5.

Confeigner and	APP NOT ALL THE CARD	
Own Trotabellariong-00		
No. of the local division of the local divis	Constructuated writer subject	
Analysis Ana	Description of the second se	
	1 ALL & CARLEY AND	
Tarias IN		

Figure 5: The Classifier Panel with further information about the TLR-generated tree.

## References

- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); *The WEKA Data Mining Software: An Update;* SIGKDD Explorations, Volume 11, Issue 1.
- [2] Weiwei Cheng, Prof. Dr. Hüllermeier, Johannes Fürnkranz, Klaus Brinker Label ranking by learning pairwise preferences, Elsevier, 2008
- [3] Weiwei Cheng, Jens Hühn, Prof. Dr. Hüllermeier Decision Tree and Instance-Based Learning for Label Ranking