Mixture-based probabilistic graphical models for the partial label ranking problem

Juan C. Alfaro[1,3], Juan A. Aledo[2,3], and José A. Gámez[1,3]
{JuanCarlos.Alfaro, JuanAngel.Aledo, Jose.Gamez}@uclm.es

[1] Departamento de Sistemas Informáticos
Universidad de Castilla-La Mancha

[2] Departamento de Matemáticas
Universidad de Castilla-La Mancha

[3] Laboratorio de Sistemas Inteligentes y Minería de Datos
Instituto de Investigación en Informática de Albacete

22nd International Conference on Intelligent Data Engineering and Automated Learning

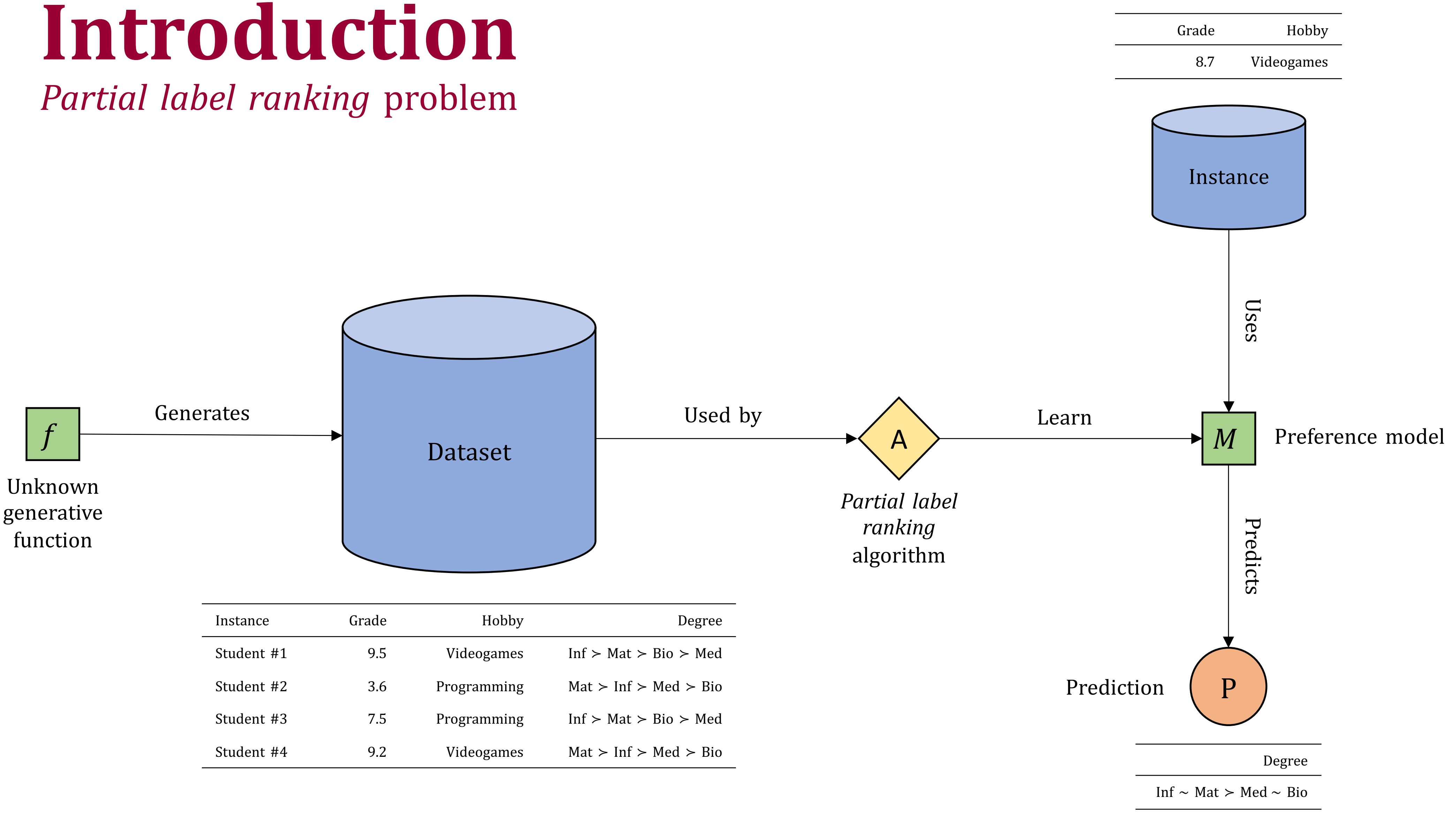
25-27 November 2021

Funding: SBPLY/17/180501/000493
Funding: PID2019–106758GB–C3
Funding: FPU18/00181

Table of contents

Introduction

Partial label ranking problem



Grade	Hobby
8.7	Videogames

Instance	Grade	Hobby	Degree
Student #1	9.5	Videogames	Inf ≻ Mat ≻ Bio ≻ Med
Student #2	3.6	Programming	Mat ≻ Inf ≻ Med ≻ Bio
Student #3	7.5	Programming	Inf ≻ Mat ≻ Bio ≻ Med
Student #4	9.2	Videogames	Mat ≻ Inf ≻ Med ≻ Bio

Degree
Inf ∼ Mat ≻ Med ∼ Bio

Introduction

Methods

- **Adaptation methods**
  - *Instance based partial label ranking*
  - *Partial label ranking trees*
    - ✓ Disagreements
    - ✓ Distance
    - ✓ Entropy
    - ✓ Gini
- **Ensemble methods**
  - *Bootstrap aggregating*
  - *Random forests*

Background

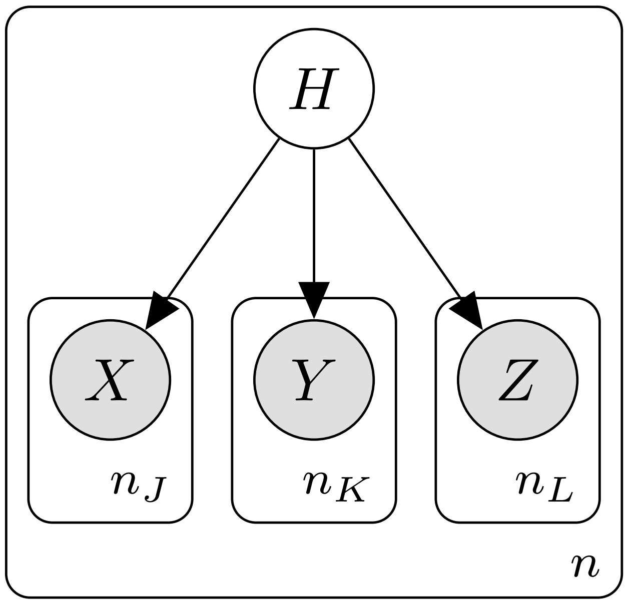
Rank aggregation problem

- A ***ranking*** represents a **precedence relation** among a set of *items*

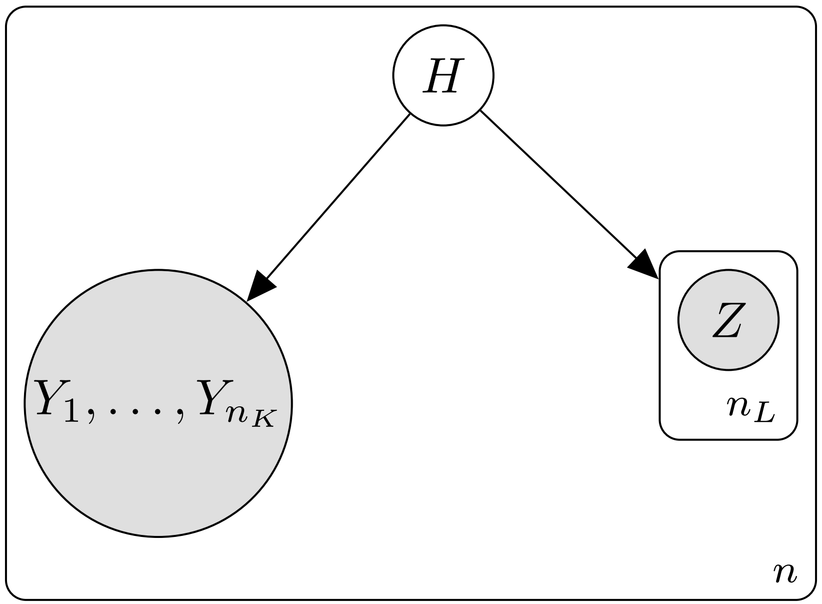
Ranking	Problem	Algorithm
Without ties	*Kemeny ranking problem*	*Borda count* algorithm
With ties	*Optimal bucket order problem*	*Bucket pivot* algorithm

Mixture models

Structure



Hidden naive bayes



Gaussian mixture semi naive bayes

Mixture models

Estimation

- ***E step***: Under the assumption that the **parameters** are **known**, we **compute** the **probability** of an **instance** being in a **mixture**
- ***M step***: Under the assumption that the **probabilities** of **belonging** to each **mixture** for all examples are **known**, the **parameters** of the model are **estimated** using **maximum likelihood estimation weighting** each **instance** by the **probability** of it being in the mixture
- **Stopping condition:** We use the ***log-likelihood*** of the model given the data with a **convergence value** of $\boldsymbol{\alpha = 0.001}$ **or** $\boldsymbol{\beta = 100}$ **máximum iterations**

Mixture models

Learning

1. We **divide** the **dataset** in **training** $\boldsymbol{Tr}$ and **validation** $\boldsymbol{Tv}$
2. We **evaluate** the **model** using $\boldsymbol{r_H = 2^1, \dots, 2^{10}}$
3. We **select** the **best** value $\boldsymbol{r'_H}$ according to $\boldsymbol{\tau_X^{Tv}}$
4. We **apply** a **binary search** in the **range** $\boldsymbol{\left[\frac{r'_H}{2}, r'_H\right]}$
5. We **select** the **best** value $\boldsymbol{r^*_H}$ according to $\boldsymbol{\tau_X^{Tv}}$
6. We **train** the **model** with the **dataset** using $\boldsymbol{r^*_H}$

Mixture models

Inference

1. We **obtain** the ***a-posteriori probability*** for the **objetive variables**
2. We **compute** the **pair order matrix** for the **input instance**
3. We **solve** the **optimal bucket order problem** to **obtain** the **output ranking**

Experiments

Datasets

Dataset	#instances	#attributes	#labels	#rankings	#buckets
authorship	841	70	4	47	3.063
blocks	5472	10	5	116	2.337
breast	109	9	6	62	3.925
ecoli	336	7	8	179	4.140
glass	214	9	6	105	4.089
iris	150	4	3	7	2.380
letter	20000	16	26	15014	7.033
libras	360	90	15	356	6.889
pendigits	10992	16	10	3327	3.397

Experiments

Algorithms

- ***Instance Based Partial Label Ranking***
- ***Partial Label Ranking Trees***
- ***Hidden Naive Bayes***
- ***Gaussian Mixture Semi Naive Bayes***

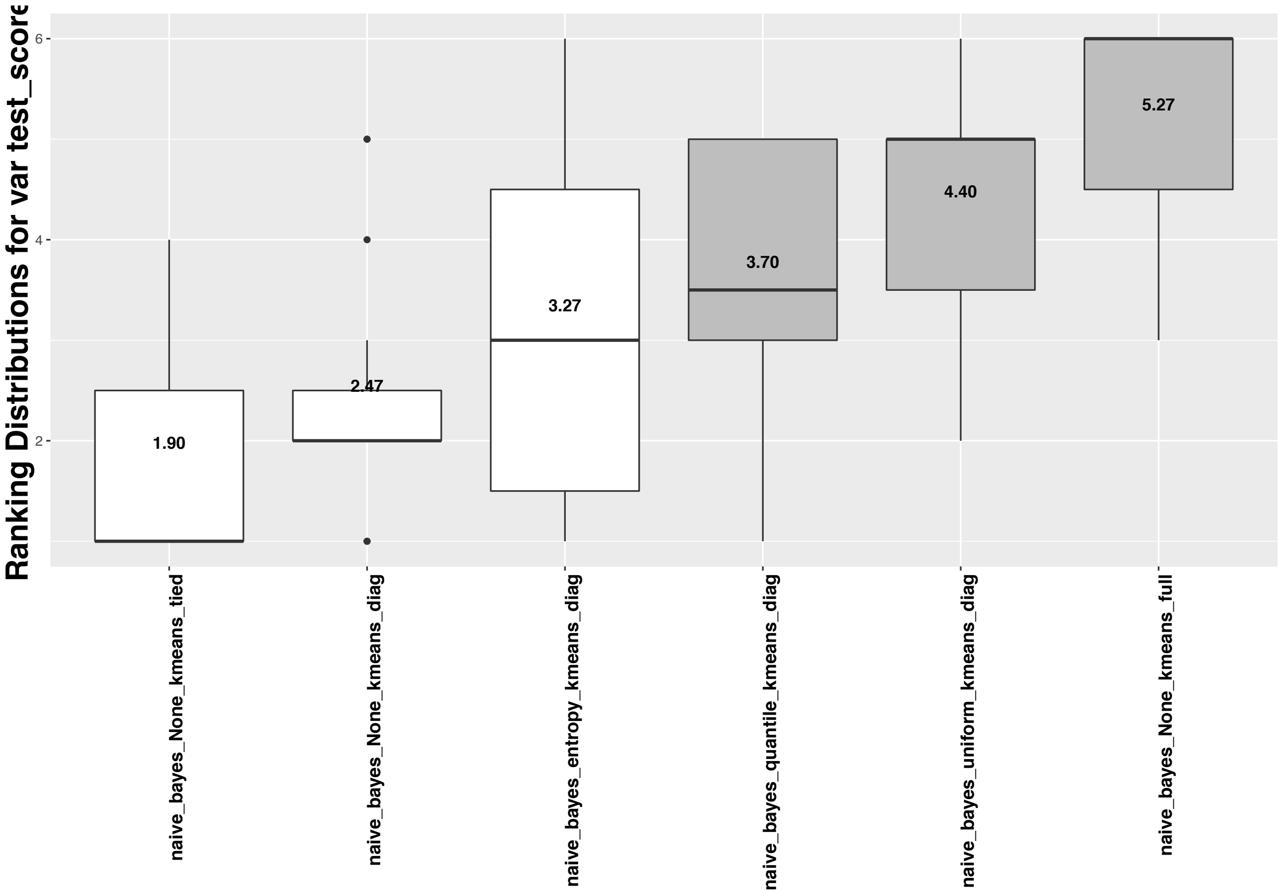
Experiments

Methodology

- The algorithms were evaluated with a $\mathbf{5 \times 10 - cv}$
- The accuracy was measured with the $\boldsymbol{\tau_X}$ **rank correlation coefficient**
  1. *Friedman test*
  2. *Post-hoc test*
- The **training and validation time** was measured (in seconds)

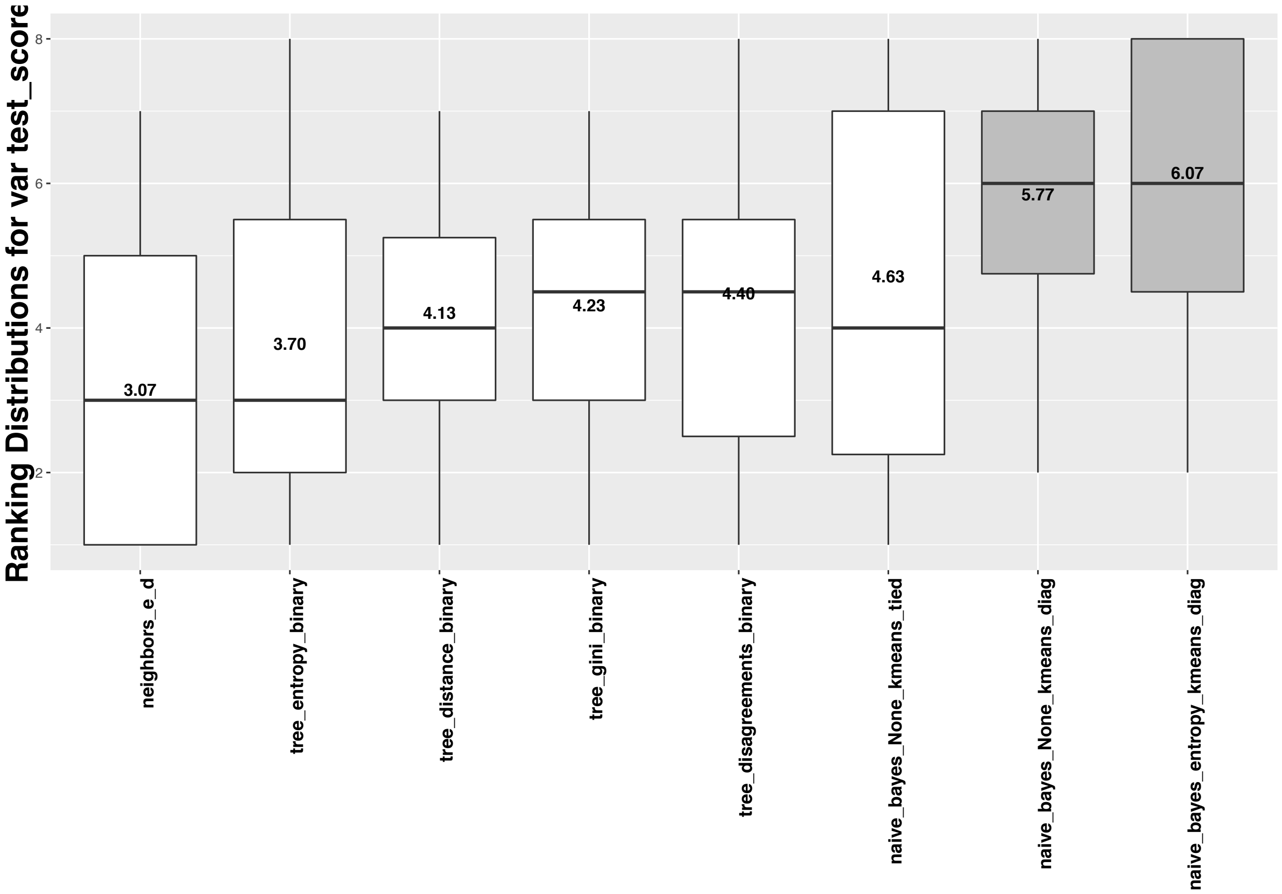
Experiments

Accuracy



Experiments

Accuracy



Experimental evaluation

Number of mixtures

Dataset	HNB-PLR-G	HNB-PLR-F	HNB-PLR-W	HNB-PLR-E	GMSNB-PLR-F	GMSNB-PLR-T
authorship	36.340 ± 36.988	24.58 ± 18.377	31.380 ± 21.813	40.840 ± 52.281	3.020 ± 0.141	35.420 ± 41.952
blocks	167.440 ± 76.169	238.400 ± 168.220	81.300 ± 33.121	**341.660 ± 147.979**	68.520 ± 23.693	213.200 ± 97.692
breast	15.960 ± 7.284	29.120 ± 19.157	19.800 ± 19.078	17.720 ± 12.795	4.520 ± 2.288	20.320 ± 8.110
ecoli	31.000 ± 14.321	25.820 ± 21.930	31.140 ± 26.869	39.900 ± 20.928	12.920 ± 6.552	47.040 ± 27.871
glass	17.220 ± 9.951	66.020 ± 32.922	27.920 ± 23.357	45.520 ± 23.603	5.180 ± 2.164	39.760 ± 16.577
iris	17.020 ± 15.946	34.820 ± 20.457	24.180 ± 17.253	9.560 ± 4.739	7.960 ± 4.000	32.320 ± 22.709
libras	47.660 ± 12.967	40.940 ± 14.621	43.960 ± 13.425	121.760 ± 38.154	**218.080 ± 39.608**	56.200 ± 10.900
pendigits	**388.800 ± 108.298**	204.680 ± 67.601	266.480 ± 95.088	261.520 ± 98.865	93.800 ± 27.355	**405.840 ± 102.542**
satimage	326.660 ± 101.758	283.660 ± 96.820	198.720 ± 108.040	272.060 ± 108.377	29.620 ± 14.246	392.460 ± 110.909
segment	140.400 ± 56.351	202.580 ± 158.267	196.300 ± 154.706	230.320 ± 141.072	42.680 ± 21.920	337.380 ± 121.548
vehicle	73.320 ± 35.361	**292.600 ± 151.414**	**346.300 ± 144.20**	172.420 ± 126.264	12.260 ± 2.448	66.480 ± 60.716
vowel	75.320 ± 26.250	169.260 ± 55.564	186.260 ± 49.278	95.640 ± 42.740	7.640 ± 2.884	174.580 ± 52.597
wine	6.700 ± 9.033	11.980 ± 15.213	17.120 ± 19.256	24.960 ± 23.206	3.800 ± 1.030	14.480 ± 17.117
yeast	103.500 ± 56.536	46.000 ± 18.553	127.660 ± 97.812	159.040 ± 113.482	30.300 ± 16.656	219.880 ± 93.535

Conclusions

- The ***gaussian mixture semi naive bayes*** algorithm is **competitive** in **accuracy** with respect to the ***instance based partial label ranking*** and ***partial label ranking trees*** methods
- The ***gaussian mixture semi naive bayes*** algorithm is **faster** during the **inference** phase than the ***instance based partial label ranking*** method
- The gaussian and entropy ***hidden naive bayes*** algorithms are **competitive** in **accuracy** with respect to the ***gaussian mixture semi naive bayes*** method

Future research lines

- We plan to **allow** training **datasets** labeled with (**possibly incomplete**) **partial rankings**
- We plan to **adapt** and **use *multilabel* algorithms** to the partial label ranking problem
- We plan to **reduce** the **problem** using **clustering techniques**

Mixture-based probabilistic graphical models for the *partial label ranking* problem

Juan C. Alfaro[1,3], Juan A. Aledo[2,3], and José A. Gámez[1,3]
{JuanCarlos.Alfaro, JuanAngel.Aledo, Jose.Gamez}@uclm.es

[1] Departamento de Sistemas Informáticos
Universidad de Castilla-La Mancha

[2] Departamento de Matemáticas
Universidad de Castilla-La Mancha

[3] Laboratorio de Sistemas Inteligentes y Minería de Datos
Instituto de Investigación en Informática de Albacete

22nd International Conference on Intelligent Data Engineering and Automated Learning

25-27 November 2021

Castilla-La Mancha
Consejería de Educación, Cultura y Deportes
MINISTERIO DE CIENCIA E INNOVACIÓN
UNIÓN EUROPEA
Fondo Social Europeo
El FSE invierte en tu futuro
AGENCIA ESTATAL DE INVESTIGACIÓN

Funding: SBPLY/17/180501/000493
Funding: PID2019–106758GB–C3
Funding: FPU18/0018