A Novel Online Stacked Ensemble for Multi-Label Stream Classification

Alican Büyükçakır, Hamed Bonab, Fazli Can

Alican Büyükçakır

alicanbuyukcakir@bilkent.edu.tr

http://abuyukcakir.github.io

Bilkent Information Retrieval Group – Computer Engineering Department
Bilkent University

ACM CIKM 2018, Torino, Italy
22-26 Oct 2018

Outline

Introduction
- Motivation
- Problem Definition

Related Work
- Related Work in MLC
- Related Work in Ensembles of MLSC

GOOWE-ML
- The Idea
- Ensemble Maintenance
- Weight Assignment and Update
- Multi-label Prediction

Experiments and Results

Discussion and Conclusion

Outline

Introduction
- Motivation
- Problem Definition

Related Work
- Related Work in MLC
- Related Work in Ensembles of MLSC

GOOWE-ML
- The Idea
- Ensemble Maintenance
- Weight Assignment and Update
- Multi-label Prediction

Experiments and Results

Discussion and Conclusion

Motivation

"A novel online stacked ensemble for multi-label stream classification."

Motivation

"A novel online stacked ensemble for multi-label stream classification."

- online, stream: Can see once, time and memory limitations, concept drifts

Motivation

"A novel online stacked ensemble for multi-label stream classification."

- online, stream: Can see once, time and memory limitations, concept drifts
- multi-label: Classification into a subset of L labels– 2^L combinations.
  - text tagging, gene function prediction, movie into genre classification...

Motivation

"A novel online stacked ensemble for multi-label stream classification."

- online, stream: Can see once, time and memory limitations, concept drifts
- multi-label: Classification into a subset of L labels– 2^L combinations.
  - text tagging, gene function prediction, movie into genre classification...
- stacked ensemble: Stacking using a geometric interpretation of label spaces.

Problem Definition: Multi-label Stream Classification

The problem is Multi-label Stream Classification (MLSC), involving complicacies of multi-label learning as well as data streams.

Problem Definition: Multi-label Stream Classification

The problem is Multi-label Stream Classification (MLSC), involving complicacies of multi-label learning as well as data streams.

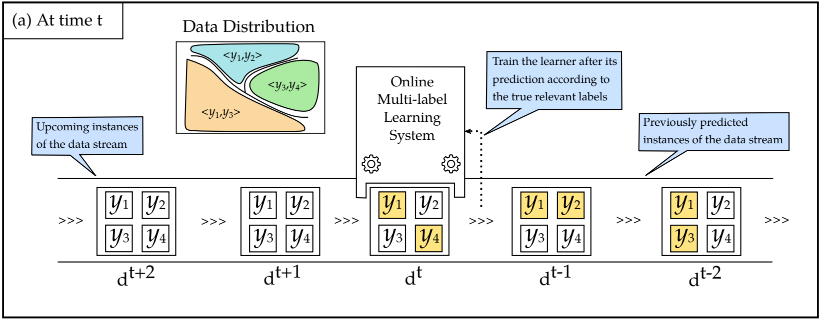

Figure: MLSC task with $L = 4$. Labels predicted as relevant are filled with yellow. Also, see ITTT (interleaved-test-then-train).

Problem Definition: Multi-label Stream Classification

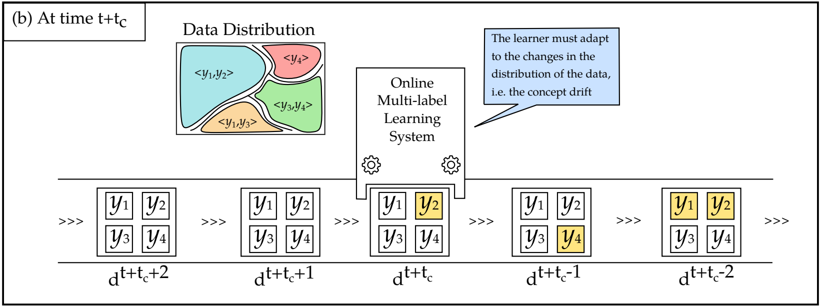

Figure: t_c units of time later, a concept drift happens. Now, the learner must modify itself according to the changes in the distribution of the data.

Outline

Introduction
- Motivation
- Problem Definition

Related Work
- Related Work in MLC
- Related Work in Ensembles of MLSC

GOOWE-ML
- The Idea
- Ensemble Maintenance
- Weight Assignment and Update
- Multi-label Prediction

Experiments and Results

Discussion and Conclusion

Related Work in MLC

Two main ways [Zhang and Zhou, 2014] to tackle multi-label problems:

1. Problem Transformation
   - Binary Relevance [Tsoumakas and Katakis, 2006]
   - Classifier Chains [Read et al., 2009]
   - Pruned Sets [Read et al., 2008]
   - Pairwise Methods [Fürnkranz et al., 2008]
2. Algorithm Adaptation
   - Decision Trees [Clare and King, 2001, Read et al., 2012]
   - ML-KNN [Zhang and Zhou, 2007]
   - Trees + Perceptrons [Osojnik et al., 2017]
   - Rule learners [Sousa and Gama, 2018]

Problem Transformation

1. **Binary Relevance**: Treat multi-label problem with L labels as L different single-label problems. Fails to capture dependencies among labels.
2. **Classifier Chains**: Randomly permute the labels and feed outputs of one label to the next ones as features.
3. **Pruned Sets**: Work on the most common subset of labels as if they are individual labels.
4. **Pairwise Methods**: Generate classifiers for each pairs of labels. Complexity is quadratic in L. Generally in the Label Ranking context.

Transform the data so that it fits your algorithm.

After PT, use Hoeffding Trees [Domingos and Hulten, 2000] to classify instances in the data stream.

Algorithm Adaptation

1. **Decision Tree-based**: Change the split criterion for different decision tree models. E.g. use Multi-label entropy.
2. **KNN-based**: Look at the nearest neighbors in the feature space for multi-label prediction.
3. **Rule Learners**: Establish association rules between features and labels such as:

$$(x_3 < 5) \wedge (x_2 > 2) \implies (y_1 = 1)$$

then combine the outputs of the rules for multi-label prediction.

Change your algorithm so that it fits your data.

Related Work in Ensembles of MLSC

1. Static Weighting + component-agnostic: Online Bagging [Oza, 2005]
    - EBR, ECC, EPS, E_BRT, E_BMT, ML-Random Rules

Related Work in Ensembles of MLSC

1. Static Weighting + component-agnostic: Online Bagging [Oza, 2005]
   - EBR, ECC, EPS, E_BRT, E_BMT, ML-Random Rules
2. Static Weighting + component-agnostic + explicit change detector: ADWIN Bagging [Bifet and Gavalda, 2007]
   - E_aBR, E_aCC, E_aPS, $E_a HT_{PS}$

Related Work in Ensembles of MLSC

1. Static Weighting + component-agnostic: Online Bagging [Oza, 2005]
   - EBR, ECC, EPS, E_BRT, E_BMT, ML-Random Rules
2. Static Weighting + component-agnostic + explicit change detector: ADWIN Bagging [Bifet and Gavalda, 2007]
   - E_aBR, E_aCC, E_aPS, E_aHT$_{PS}$
3. Dynamic weighting + component-sensitive:
   - MW, SMART, SWMEC

Related Work in Ensembles of MLSC

1. Static Weighting + component-agnostic: Online Bagging [Oza, 2005]
   - EBR, ECC, EPS, E_BRT, E_BMT, ML-Random Rules
2. Static Weighting + component-agnostic + explicit change detector: ADWIN Bagging [Bifet and Gavalda, 2007]
   - E_aBR, E_aCC, E_aPS, E_aHT$_{PS}$
3. Dynamic weighting + component-sensitive:
   - MW, SMART, SWMEC
4. Dynamic weighting + component-agnostic: **GOOWE-ML**

Bilkent University

Outline

Introduction
- Motivation
- Problem Definition

Related Work
- Related Work in MLC
- Related Work in Ensembles of MLSC

GOOWE-ML
- The Idea
- Ensemble Maintenance
- Weight Assignment and Update
- Multi-label Prediction

Experiments and Results

Discussion and Conclusion

GOOWE-ML

Geometrically Optimum Online Weighted Ensemble for Multi-Label Classification.

- Batch-incremental
- Dynamic sized (evolving)
- Weighting of the components..?

The Idea: Origins in Data Fusion

The idea behind GOOWE-ML (and its single-label counterpart GOOWE [Bonab and Can, 2018]) is actually from the field of Data Fusion.

The Idea: Origins in Data Fusion

The idea behind GOOWE-ML (and its single-label counterpart GOOWE [Bonab and Can, 2018]) is actually from the field of Data Fusion.

[Wu and Crestani, 2015]'s work: "A geometric framework for data fusion in information retrieval".

Similarity between the problems

Finding relevant documents for a query by combining outputs of multiple information retrieval systems: very similar to MLC task.

The Idea: Origins in Data Fusion

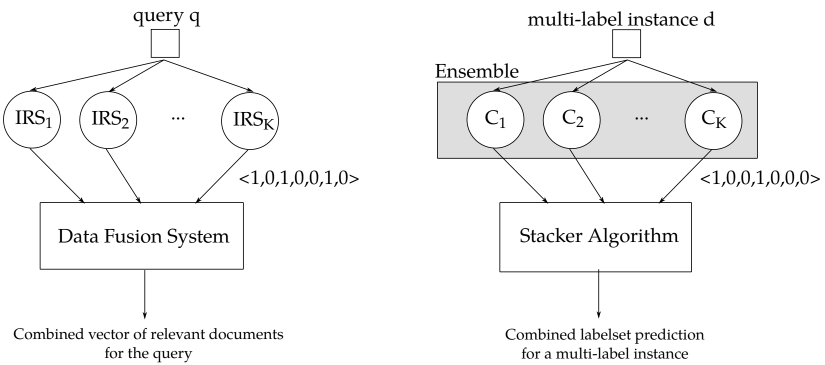

Figure: An IRS's response to a query q is a vector that is similar to an MLC system's prediction for a data instance. Data Fusion scheme for multiple IRSs is analogous to combiner algorithm for a stacked ensemble.

Ensemble Maintenance

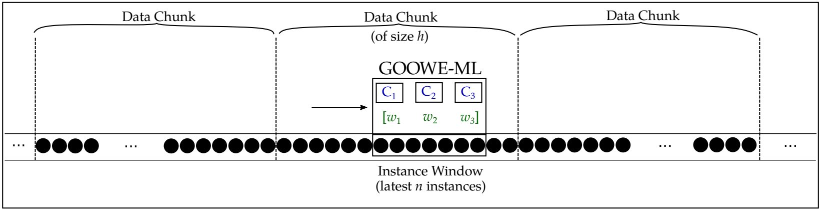

Figure: General view of GOOWE-ML.

- Train a new classifier at each Data Chunk. Growing Ensemble.
- Adjust weights at the end of each Data Chunk.
- If full, replace the component with the lowest weight.
- Prequential Evaluation using Instance Window.

Weight Assignment and Update: Label Space

In GOOWE-ML, we represent each relevance score vector in an L-dimensional space (label space).

Related

Similar idea in [Tai and Lin, 2012], *Principal Label-Space Transformation*. They used this idea to reduce the dimensionality of the label-space and reinterpreting the existing multi-label methods in this setting.

Weight Assignment and Update: Representation

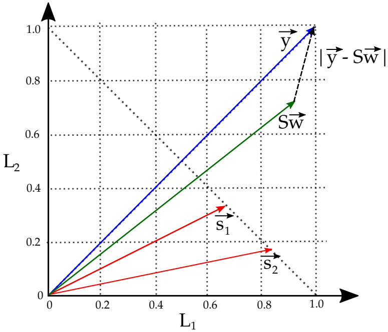

Figure: Transformation into label space in GOOWE-ML. Relevance scores of the components (red): $S_1 =< 0.65, 0.35 >$ and $S_2 =< 0.82, 0.18 >$. The optimal vector $\vec{y}$ (blue): $y =< 1, 1 >$, generated from the ground truth. Weighted prediction of the ensemble: $S\vec{w}$ (green). The distance between $\vec{y}$ and $S\vec{w}$ is minimized.

Weight Assignment: Solving Linear Least Squares

The linear least squares problem:

$$\min_{\vec{w}} ||\vec{y} - S\vec{w}||_2^2 \tag{1}$$

Weight Assignment: Solving Linear Least Squares

The linear least squares problem:

$$\min_{\vec{w}} ||\vec{y} - S\vec{w}||_2^2 \tag{1}$$

The objective function:

$$f(W_1, W_2, .., W_K) = \sum_{i=1}^{n} \sum_{j=1}^{L} \left(\sum_{k=1}^{K} (W_k S_{kj}^i - y_j^i) \right)^2 \tag{2}$$

Weight Assignment: Solving Linear Least Squares

Setting the gradient $\nabla f = 0$ and redistributing the terms, we get:

$$\sum_{k=1}^{K} W_k \left(\sum_{i=1}^{n} \sum_{j=1}^{L} S_{qj}^{i} S_{kj}^{i} \right) = \sum_{i=1}^{n} \sum_{j=1}^{L} y_j^i S_{qj}^{i} \tag{3}$$

Weight Assignment: Solving Linear Least Squares

Setting the gradient $\nabla f = 0$ and redistributing the terms, we get:

$$\sum_{k=1}^{K} W_k \left(\sum_{i=1}^{n} \sum_{j=1}^{L} S_{qj}^{i} S_{kj}^{i} \right) = \sum_{i=1}^{n} \sum_{j=1}^{L} y_j^i S_{qj}^{i} \tag{3}$$

Let's rewrite this as a matrix-vector multiplication. This is eqv. to $Aw = d$ where A is a matrix with elements:

Weight Assignment: Solving Linear Least Squares

Setting the gradient $\nabla f = 0$ and redistributing the terms, we get:

$$\sum_{k=1}^{K} W_k \left(\sum_{i=1}^{n} \sum_{j=1}^{L} S_{qj}^{i} S_{kj}^{i} \right) = \sum_{i=1}^{n} \sum_{j=1}^{L} y_j^i S_{qj}^{i} \tag{3}$$

Let's rewrite this as a matrix-vector multiplication. This is eqv. to $Aw = d$ where A is a matrix with elements:

$$a_{qk}^{i} = \sum_{i=1}^{n} \sum_{j=1}^{L} S_{qj}^{i} S_{kj}^{i} \qquad (1 \leq q, k \leq K) \tag{4}$$

Weight Assignment: Solving Linear Least Squares

Setting the gradient $\nabla f = 0$ and redistributing the terms, we get:

$$\sum_{k=1}^{K} W_k \left(\sum_{i=1}^{n} \sum_{j=1}^{L} S_{qj}^{i} S_{kj}^{i} \right) = \sum_{i=1}^{n} \sum_{j=1}^{L} y_j^i S_{qj}^{i} \tag{3}$$

Let's rewrite this as a matrix-vector multiplication. This is eqv. to $Aw = d$ where A is a matrix with elements:

$$a_{qk}^{i} = \sum_{i=1}^{n} \sum_{j=1}^{L} S_{qj}^{i} S_{kj}^{i} \qquad (1 \leq q, k \leq K) \tag{4}$$

and d is a vector with elements:

$$d_q^i = \sum_{i=1}^{n} \sum_{j=1}^{L} y_j^i S_{qj}^{i} \qquad (1 \leq q \leq K) \tag{5}$$

Weight Assignment and Update

Therefore, geometrically-optimum weight assignment is as follows:

- At each instance, save the relevance scores of each classifier and the ground truth for each multi-label instance.
- When the current data chunk is filled,
  1. Train the new and the existing classifiers,
  2. Then, populate the matrix A and vector d.
  3. Solve the linear system $Aw = d$ to get the optimal weight vector.
  4. Replace the component with the lowest weight.

Multi-label Prediction

Normalized and *thresholded* weighted sum of ensemble's components' relevance scores.

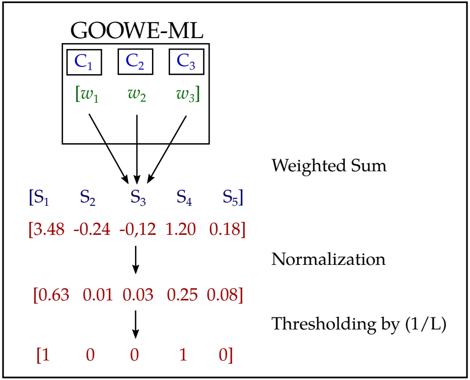

Figure: Multi-label prediction of GOOWE-ML. Example with $L = 5$.

Outline

Introduction
- Motivation
- Problem Definition

Related Work
- Related Work in MLC
- Related Work in Ensembles of MLSC

GOOWE-ML
- The Idea
- Ensemble Maintenance
- Weight Assignment and Update
- Multi-label Prediction

Experiments and Results

Discussion and Conclusion

Setup

- Generate 4 different types of GOOWE-ML-based classifiers: **GOBR**, **GOCC**, **GOPS**, **GORT** using different problem transformations. As a base classifier, use Hoeffding Trees.
- Compare with the following baselines [Read et al., 2012], [Osojnik et al., 2017]: EBR, ECC, EPS, E_BRT, E_aBR, E_aCC, E_aPS.
- Experiment on 7 datasets.
- Friedman test with Nemenyi post-hoc analysis for statistical significance [Demšar, 2006].

Prequential Evaluation

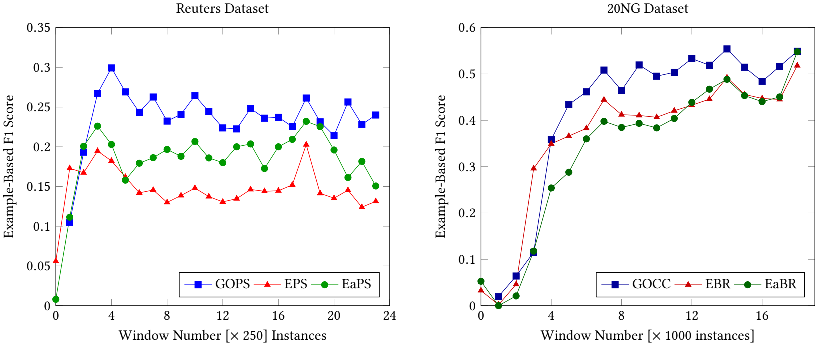

Figure: Prequential Evaluation of Models: Example-Based F1 Score for Reuters and 20NG datasets.

Results

Table: Experimental Results: Example-based F1 Score

	20NG	Yeast	Ohsumed	Slashdot	Reuters	IMDB	TMC7	
(a) Example-Based F1 Score ($F1_{ex}$) $\uparrow$								**Avg. Rank**
GOBR	0.364	0.650	0.307	0.189	0.076	0.283	0.623	4.00
GOCC	**0.442**	**0.652**	**0.352**	0.028	0.145	0.221	**0.668**	**2.57**
GOPS	0.224	0.644	0.331	**0.405**	**0.252**	**0.333**	0.485	3.00
GOBRT	0.196	0.607	0.297	0.189	0.078	0.283	0.452	5.71
EBR	0.365	0.638	0.23	0.023	0.106	0.075	0.654	4.71
ECC	0.349	0.632	0.217	0.020	0.098	0.016	0.643	6.43
EPS	0.096	0.584	0.213	0.269	0.148	0.133	0.330	6.71
EBRT	0.100	0.509	0.056	0.001	0.000	0.001	0.008	10.57
EaBR	0.341	0.638	0.202	0.018	0.059	0.031	0.661	6.57
EaCC	0.156	0.633	0.005	0.020	0.004	0.001	0.646	8.14
EaPS	0.109	0.578	0.200	0.258	0.183	0.104	0.384	6.85

Results II

Table: Experimental Results: Example-based Accuracy

	20NG	Yeast	Ohsumed	Slashdot	Reuters	IMDB	TMC7	
(d) Example-Based Accuracy (Acc_{ex}) $\uparrow$								**Avg. Rank**
GOBR	0.239	0.508	0.184	0.106	0.040	0.164	0.457	4.57
GOCC	**0.391**	**0.509**	**0.277**	0.025	0.120	0.138	0.515	**3.00**
GOPS	0.137	0.504	0.211	**0.299**	0.160	**0.204**	0.327	3.29
GOBRT	0.115	0.454	0.178	0.107	0.040	0.164	0.298	6.71
EBR	0.352	0.502	0.191	0.020	0.098	0.055	0.520	4.29
ECC	0.337	0.493	0.180	0.018	0.093	0.012	0.511	6.14
EPS	0.094	0.460	0.180	0.260	0.143	0.105	0.246	6.29
EBRT	0.100	0.372	0.049	0.001	0.000	0.001	0.007	10.57
EaBR	0.330	0.502	0.169	0.016	0.056	0.024	**0.529**	6.14
EaCC	0.152	0.495	0.004	0.018	0.004	0.001	0.516	7.71
EaPS	0.108	0.455	0.170	0.250	**0.179**	0.083	0.290	6.43

Results III

Table: Experimental Results: Micro-Averaged F1 Score

	20NG	Yeast	Ohsumed	Slashdot	Reuters	IMDB	TMC7	
(b) Micro-Averaged F1 Score ($F1_{micro}$) ↑								**Avg. Rank**
GOBR	0.237	0.638	0.291	0.187	0.076	0.276	0.584	4.86
GOCC	**0.516**	**0.640**	**0.410**	0.050	0.196	0.228	0.634	**2.71**
GOPS	0.206	0.629	0.298	**0.315**	**0.210**	**0.314**	0.447	3.43
GOBRT	0.153	0.598	0.270	0.187	0.077	0.277	0.439	6.57
EBR	0.499	0.631	0.294	0.041	0.141	0.099	0.638	4.29
ECC	0.486	0.625	0.280	0.037	0.134	0.025	0.631	6.14
EPS	0.115	0.584	0.216	0.286	0.162	0.138	0.342	7.00
EBRT	0.174	0.519	0.076	0.001	0.000	0.001	0.008	10.58
EaBR	0.477	0.632	0.266	0.033	0.081	0.041	**0.640**	5.71
EaCC	0.262	0.627	0.007	0.037	0.007	0.002	0.632	7.71
EaPS	0.180	0.580	0.205	0.278	0.200	0.118	0.378	6.71

Results IV

Table: Experimental Results: Hamming Score — what happened to GOOWE-ML-based ensembles?

	20NG	Yeast	Ohsumed	Slashdot	Reuters	IMDB	TMC7	
(c) Hamming Score ↑								**Avg. Rank**
GOBR	0.749	0.769	0.738	0.625	0.707	0.727	0.886	9.86
GOCC	0.952	0.771	0.932	0.946	0.984	0.887	0.916	5.57
GOPS	0.769	0.754	0.830	0.872	0.956	0.836	0.854	9.29
GOBRT	0.624	0.716	0.730	0.644	0.720	0.732	0.815	10.57
EBR	**0.961**	0.786	**0.936**	0.946	**0.986**	0.925	0.934	2.14
ECC	**0.961**	0.786	**0.936**	**0.947**	**0.986**	0.928	0.934	**<u>1.57</u>**
EPS	0.924	0.764	0.918	0.937	0.985	0.919	0.911	7.29
EBRT	0.952	0.773	0.930	0.946	**0.986**	**0.929**	0.902	4.00
EaBR	**0.961**	0.786	0.935	0.946	**0.986**	0.928	**0.935**	2.00
EaCC	0.955	**0.787**	0.928	**0.947**	**0.986**	**0.929**	0.934	2.29
EaPS	0.950	0.767	0.918	0.937	0.985	0.924	0.913	6.71

Investigating low Hamming Scores of GOOWE-ML-based models

We found out that low Hamming Scores are related to Precision vs Recall of the models.

Table: Micro Precision (Prec) vs Recall (Rec), and Their Effect on Hamming Score (HS)

	20NG			**Ohsumed**			**Reuters**		
	Prec	Rec	HS	Prec	Rec	HS	Prec	Rec	HS
GOBR	0.140	**0.757**	0.749	0.181	**0.743**	0.738	0.040	**0.848**	0.707
GOPS	0.125	**0.580**	0.769	0.212	**0.500**	0.830	0.140	**0.418**	0.956
EBR	**0.753**	0.373	0.961	**0.713**	0.185	0.936	**0.510**	0.082	0.986
EPS	**0.142**	0.096	0.924	**0.348**	0.157	0.918	**0.361**	0.105	0.985

Why?

Investigating low Hamming Scores

Why? The answer has two components:

1. **The nature of the metric itself**.

$$\text{Hamming Score} = \frac{1}{LN}\sum_{i=1}^{N}\sum_{j=1}^{L}[\![y_j^i = \hat{y}_j^i]\!]$$

It incorporates TN's. Why is that a problem?

Investigating low Hamming Scores

Why? The answer has two components:

1. **The nature of the metric itself**.

$$\text{Hamming Score} = \frac{1}{LN}\sum_{i=1}^{N}\sum_{j=1}^{L}[\![y_j^i = \hat{y}_j^i]\!]$$

   It incorporates TN's. Why is that a problem?

2. **Label density of the datasets**. When there are many labels with few of them relevant on average (e.g. tagging tasks), then the contribution of TNs is very high.

Investigating low Hamming Scores

Why? The answer has two components:

1. **The nature of the metric itself**.

$$\text{Hamming Score} = \frac{1}{LN}\sum_{i=1}^{N}\sum_{j=1}^{L}[\![y_j^i = \hat{y}_j^i]\!]$$

   It incorporates TN's. Why is that a problem?

2. **Label density of the datasets**. When there are many labels with few of them relevant on average (e.g. tagging tasks), then the contribution of TNs is very high.

In datasets with low label density, Hamming Score is misleading as it is ALWAYS very high. Retrieval-based metrics should be preferred.

Outline

Introduction
- Motivation
- Problem Definition

Related Work
- Related Work in MLC
- Related Work in Ensembles of MLSC

GOOWE-ML
- The Idea
- Ensemble Maintenance
- Weight Assignment and Update
- Multi-label Prediction

Experiments and Results

Discussion and Conclusion

Discussion and Conclusion

- We proposed a novel ensembling scheme for multi-label learning task in data streams.
- With a good ensemble maintenance strategy, a geometric stacking scheme and optimal weight assignment, we improved the performance of existing multi-label methods. Reached state-of-the-art in ensemble models in multi-label streams.

Discussion and Conclusion

- We proposed a novel ensembling scheme for multi-label learning task in data streams.
- With a good ensemble maintenance strategy, a geometric stacking scheme and optimal weight assignment, we improved the performance of existing multi-label methods. Reached state-of-the-art in ensemble models in multi-label streams.
- Details on Time and Memory Consumption of each model is available in the paper.
- Testing of Statistical Significance using Nemenyi Critical Distance Diagrams is also available in the paper.

Discussion and Conclusion

- We proposed a novel ensembling scheme for multi-label learning task in data streams.
- With a good ensemble maintenance strategy, a geometric stacking scheme and optimal weight assignment, we improved the performance of existing multi-label methods. Reached state-of-the-art in ensemble models in multi-label streams.
- Details on Time and Memory Consumption of each model is available in the paper.
- Testing of Statistical Significance using Nemenyi Critical Distance Diagrams is also available in the paper.
- We demonstrated how *Hamming Score can be misleading in datasets with very sparse labelsets*. Perhaps, the papers in the field where Hamming Score-related reward functions are optimized should be reexamined.

Thanks, and Q&A

Thanks for your patience.

Thanks to ACM SIGIR for the Student Travel Grant.

Any questions?

References I

Bifet, A. and Gavalda, R. (2007).
Learning from time-changing data with adaptive windowing.
In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 443–448. SIAM.

Bonab, H. R. and Can, F. (2018).
GOOWE: Geometrically Optimum and Online-Weighted Ensemble classifier for evolving data streams.
ACM Transactions on Knowledge Discovery from Data (TKDD), 12(2):25.

Clare, A. and King, R. D. (2001).
Knowledge discovery in multi-label phenotype data.
In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 42–53. Springer.

Demšar, J. (2006).
Statistical comparisons of classifiers over multiple data sets.
Journal of Machine learning research, 7(Jan):1–30.

Domingos, P. and Hulten, G. (2000).
Mining high-speed data streams.
In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 71–80. ACM.

Fürnkranz, J., Hüllermeier, E., Mencía, E. L., and Brinker, K. (2008).
Multilabel classification via calibrated label ranking.
Machine learning, 73(2):133–153.

Osojnik, A., Panov, P., and Džeroski, S. (2017).
Multi-label classification via multi-target regression on data streams.
Machine Learning, 106(6):745–770.

References II

Oza, N. C. (2005).
Online bagging and boosting.
In *2005 IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 2340–2345. IEEE.

Read, J., Bifet, A., Holmes, G., and Pfahringer, B. (2012).
Scalable and efficient multi-label classification for evolving data streams.
Machine Learning, 88(1-2):243–272.

Read, J., Pfahringer, B., and Holmes, G. (2008).
Multi-label classification using ensembles of pruned sets.
In *The Eighth IEEE ICDM*, pages 995–1000. IEEE.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2009).
Classifier chains for multi-label classification.
In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 254–269. Springer.

Sousa, R. and Gama, J. (2018).
Multi-label classification from high-speed data streams with adaptive model rules and random rules.
Progress in Artificial Intelligence, pages 1–11.

Tai, F. and Lin, H.-T. (2012).
Multilabel classification with principal label space transformation.
Neural Computation, 24(9):2508–2542.

Tsoumakas, G. and Katakis, I. (2006).
Multi-label classification: An overview.
International Journal of Data Warehousing and Mining, 3(3).

Bilkent University

References III

Wu, S. and Crestani, F. (2015).
A geometric framework for data fusion in information retrieval.
Information Systems, 50(Supplement C):20 – 35.

Zhang, M.-L. and Zhou, Z.-H. (2007).
Ml-knn: A lazy learning approach to multi-label learning.
Pattern Recognition, 40(7):2038–2048.

Zhang, M.-L. and Zhou, Z.-H. (2014).
A review on multi-label learning algorithms.
IEEE Transactions on Knowledge and Data Engineering, 26(8):1819–1837.