Gene Function Prediction Using SVM

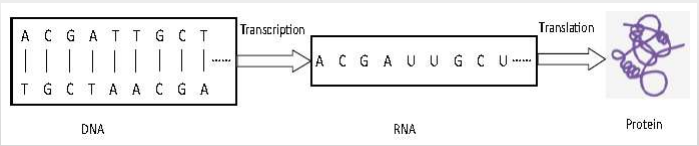
>> Multi-label Classification Based on Label Ranking and Delicate Boundary SVM

Hierarchical Multi-label Classification Based on Over-sampling and Hierarchy Constraint

Gene Function Prediction

Gene Function Prediction

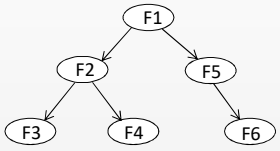
- The completion of several genome projects in the past decade has generated the full genome sequence of many organisms. **Assigning biological functions** to the sequences has become a key challenge in modern biology.
- Machine learning techniques are often used to predict gene functions from a predefined set of possible functions. Afterwards, the predictions with highest confidence can be tested in the lab.

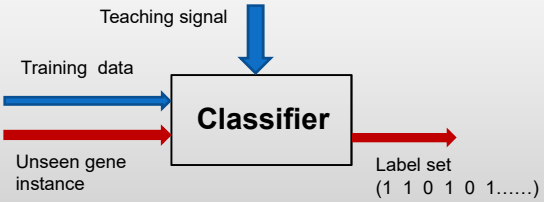


Gene Function Prediction

- Multi-label or hierarchical multi-label classification tasks.

	F 1	F 2	.....	F m
Gene 1	1	0		1
Gene 2	1	1		0
.....				
Gene n	1	1		0





Gene Function Prediction

- Data features and labels.

  Example: FunCat Yeast benchmark datasets.

Rooted tree structure

Attribute	Type	Description
aa-rat-X	real	Percentage of amino acid X in the protein
seq-len	integer	Length of the protein sequence
aa-rat-pair-X-Y	real	Percentage of the pair of amino acids X and Y consecutively in the protein
mol-wt	integer	Molecular weight of the protein
theo-pI	real	Theoretical pI (isoelectric point)
atomic-comp-X	real	Atomic composition of X where X is c (carbon), o (oxygen), n (nitrogen), s (sulphur) or h (hydrogen)
aliphatic-index	real	The aliphatic index
hydro	real	Grand average of hydrophobicity
...	...	...

Comprehensive analysis information abstracted from gene sequence

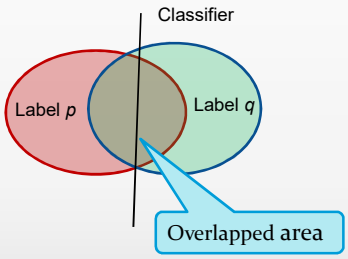
```
1 Matabolism
1.1 amino acid metabolism
1.1.3 assimilation of ammonia, metabolism of the glutamate group
1.1.3.1 metabolism of glutamine
1.1.3.1.1 biosynthesis of glutamine
1.1.3.1.2 degradation of glutamine
...
1.2 nitrogen, sulfur, and selenium metabolism
...
14 Protein fate (folding, modification, destination)
14.01 protein folding and stabilization
14.04 protein targeting, sorting and translocation
14.07 protein modification
14.07.01 modification with fatty acids
14.07.02 modification with sugar residues
14.07.02.02 N-directed glycosylation, deglycosylation
14.07.03 modification by phosphorylation
...
14.13 protein/peptide degradation
...
```

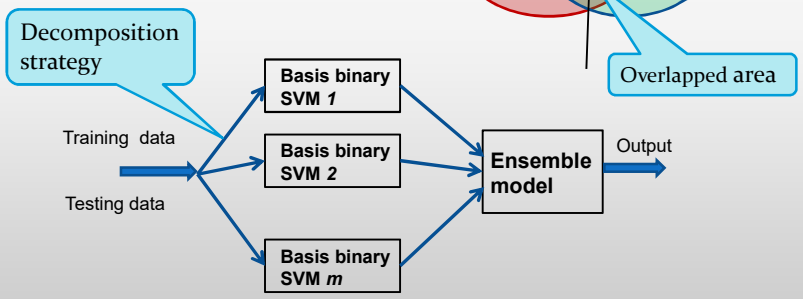
Multi-label Classification Based on Label Ranking

Introduction

- **Multi-label classification**: each sample may belong to several classes simultaneously.

	F 1	F 2	.....	F m
Gene 1	1	0		1
Gene 2	1	1		0
.....				
Gene n	1	1		0

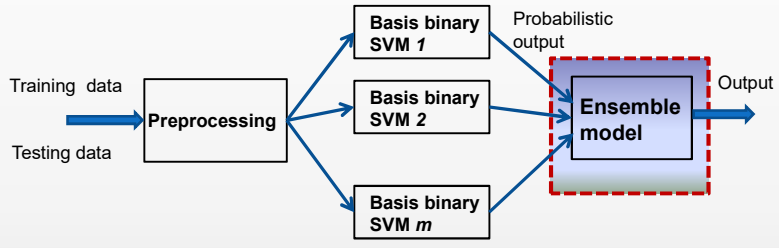


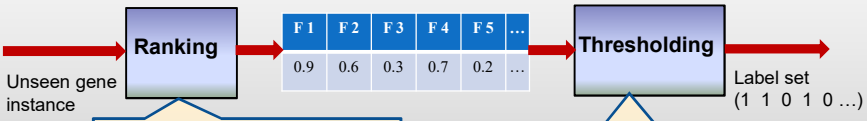


Multi-label Classification Based on Label Ranking

Motivation

- Conventional label ranking strategy.



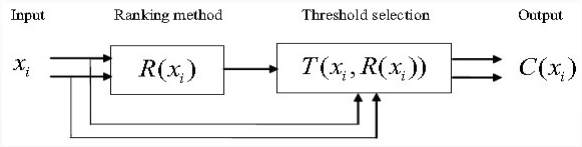


Challenge: how to deal with the overlapped samples effectively.

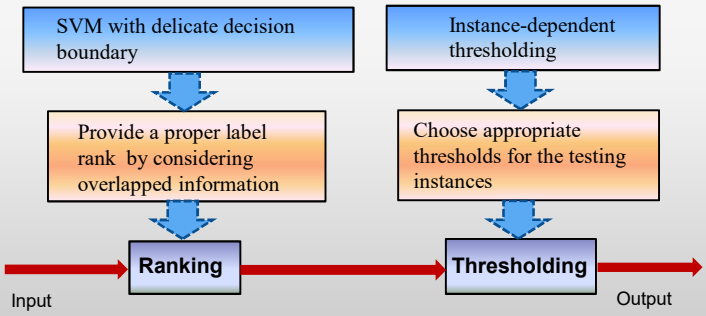
Challenge: how to select an appropriate threshold.

Multi-label Classification Based on Label Ranking

Proposed Ranking Based Multi-label Classification



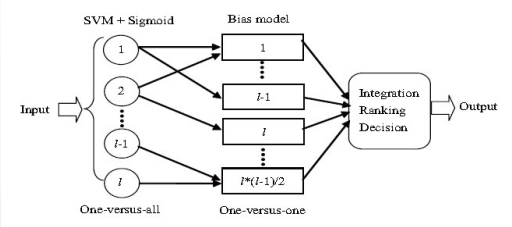
$$C(x_i) = T(x_i, R(x_i))R(x_i)$$



Multi-label Classification Based on Label Ranking

Proposed Ranking Based Multi-label Classification

- SVM with delicate decision boundary.

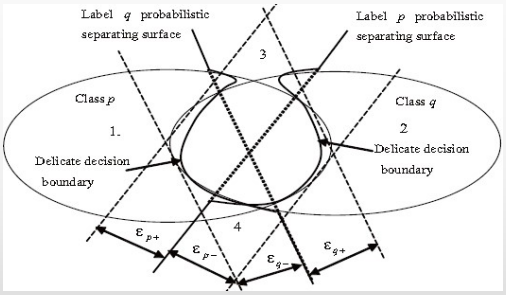


➢Probabilistic outputs for SVM: Platt's sigmoid method.
➢**The bias model** is used to correct the misclassified samples in overlapped area.

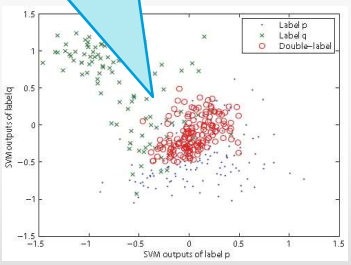
Multi-label Classification Based on Label Ranking

Proposed Ranking Based Multi-label Classification

- The bias model.



The majority of double-label samples are distributed near to two binary SVM separating surfaces simultaneously.



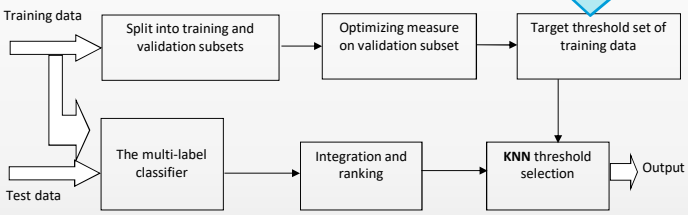
➢Four borders: the distributing range of overlapped sample space.
➢Two delicate boundaries for two label respectively.

Multi-label Classification Based on Label Ranking

Proposed Ranking Based Multi-label Classification

- Instance-dependent thresholding strategy.

The target thresholds of training samples as teachers.



➢The proposed instance-dependent thresholding strategy is considered as a function that relies on input instance and label rank.
➢A *KNN* learning model is used to realize the threshold function.

Multi-label Classification Based on Label Ranking

Experiments

- Two benchmark problems.
  - Yeast functional genomics.
  - Genbase motif-based protein classification.

Table 3.1: Characteristics of the experimental benchmark datasets

Dataset	Total labels	Total features	Avg. labels	Training/Testing set
Yeast	14	103 (Numeric)	4.25	1500/917
Genbase	27	1186 (Discrete)	1.35	463/199

Multi-label Classification Based on Label Ranking

Experiments

- Evaluation metrics.

$$HammingLoss(C,D)=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{1}{k}|y_i \Delta z_i|$$

Prediction result

$$Accuracy(C,D)=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|y_i \cap z_i|}{|y_i \cup z_i|}$$

Actual label set of test instance

$$Precision(C,D)=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|y_i \cap z_i|}{|z_i|}$$

$$Recall(C,D)=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|y_i \cap z_i|}{|y_i|}$$

Multi-label Classification Based on Label Ranking

Experiments

- The experiment results for two datasets.

Table 3.2: Experimental results for Yeast dataset.

Algorithms	Hamming loss	Accuracy	Precision	Recall
C4.5	0.259	0.423	0.561	0.593
Naive Bayes	0.301	0.421	0.610	0.531
Binary-SVM (based on RBF kernel)	0.2021	0.5302	0.5862	0.6332
Proposed (based on RBF kernel)	**0.1915**	**0.5481**	**0.6761**	**0.7106**

Table 3.3: Experimental results for Genbase dataset.

Algorithms	Hamming loss	Accuracy	Precision	Recall
C4.5	0.001	0.987	0.992	0.995
Naive Bayes	0.035	0.273	0.273	0.276
Binary-SVM (Based on linear SVM)	0.0011	0.9891	0.9933	**0.9958**
Proposed (based on linear SVM)	0.00006	0.9912	0.9947	0.9950
Proposed (based on ck-SVM)	**0.00006**	**0.9925**	**0.9968**	0.9942

Multi-label Classification Based on Label Ranking

Experiments

- Experiment results.

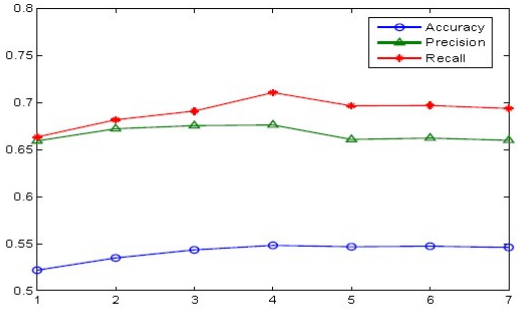


Figure 3.5: Relation between the parameter k of KNN in threshold selection and the performance of classification for Yeast dataset

Multi-label Classification Based on Label Ranking

Experiments

- Experiment results.

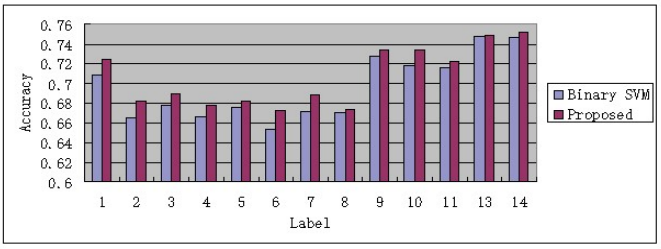
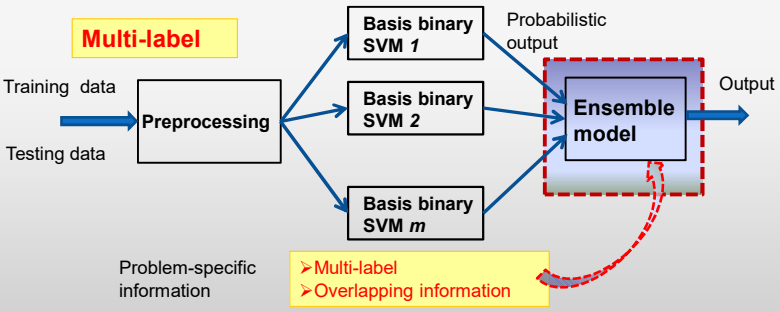


Figure 3.6: Comparing results of the bias models for 12-th label in Yeast dataset

Multi-label Classification Based on Label Ranking

Summary

1) A novel scoring method considering the information of overlapped training samples is proposed. A **bias model** with delicate decision boundaries is used to improve the classification accuracy of overlapping samples.
2) An **instance-dependent thresholding** of relating with input instance and label rank is proposed to decide the classification results.



Part 2: Gene Function Prediction

Multi-label Classification Based on Label Ranking and Delicate Boundary SVM

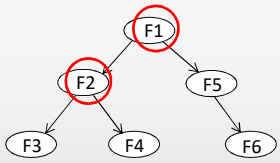
>> Hierarchical Multi-label Classification Based on Over-sampling and Hierarchy Constraint

Hierarchical Multi-label Classification (HMC)

Introduction

- **Hierarchical multi-label classification (HMC)** : an example that belongs to one class automatically belongs to all its super-classes (this is called the hierarchy constraint).

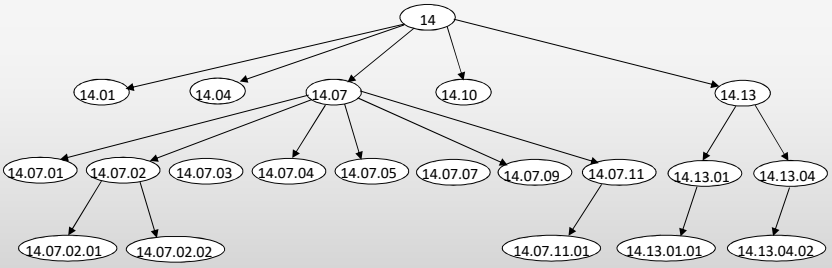
	F 1	F 2	.....	F m
Gene 1	1	0		1
Gene 2	1	1		0
.....				
Gene n	1	1		0



Hierarchical Multi-label Classification (HMC)

Introduction

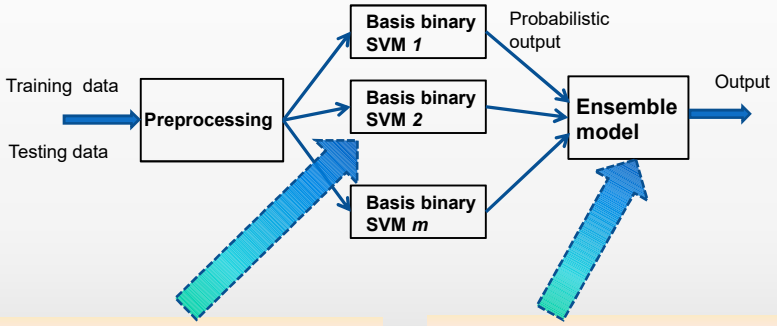
- Gene function predictions on **FunCat taxonomy** are difficult HMC tasks.
  - Hundreds of functional classes: > 400.
  - High degree imbalanced : negative >> positive.



Hierarchical Multi-label Classification (HMC)

Motivations

- **Two main challenges.**

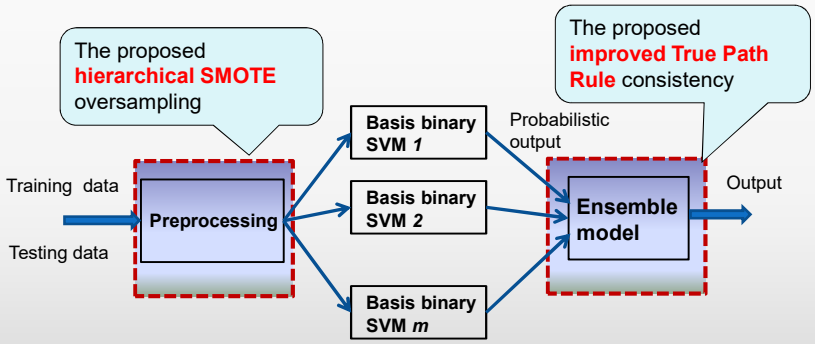


Imbalance dataset learning:
standard classifiers tend to be **overwhelmed** by the large classes and ignore the small ones.

Hierarchy constraint:
the hierarchy constraint of classes is not considered in binary classifiers.

Hierarchical Multi-label Classification (HMC)

Improved HMC Method



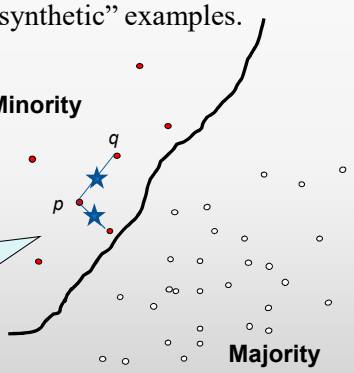
Hierarchical Multi-label Classification (HMC)

Proposed Hierarchical SMOTE

- **Imbalanced dataset learning.**

SMOTE (Synthetic Minority Over-sampling Technique) [Chawla 2002] is an over-sampling approach in which the minority class is over-sampled by creating "synthetic" examples.

The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the *k* minority class nearest neighbors.



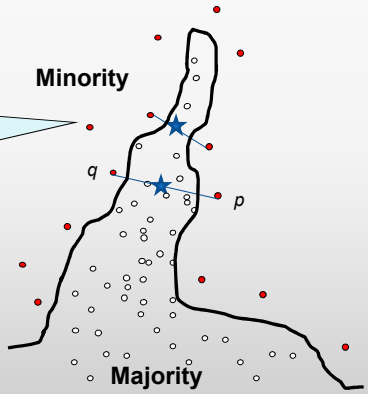
Hierarchical Multi-label Classification (HMC)

Proposed Hierarchical SMOTE

- **Drawback of SMOTE**

  The basic SMOTE method has a drawback in some special situations, it can not obtain good performance because of its assumptions about the training set.

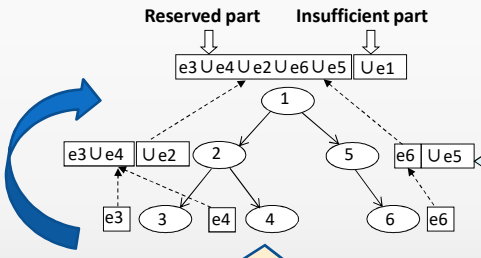
The space between two positive instances is assumed to be positive, but those assumptions may not always be true for some special distribution datasets.



Hierarchical Multi-label Classification (HMC)

Proposed Hierarchical SMOTE

- The proposed **hierarchical SMOTE**.



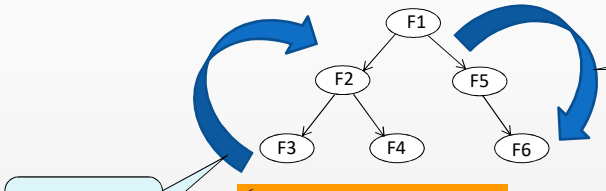
Over-sampling procedures are implemented from bottom to top across the class tree, and the "synthetic" examples created by children classes are automatically reserved and regarded as the "synthetic" examples of their parent classes.

Two advantages:
- Can remarkably reduce the over-sampling operations;
- Can solve the drawback of basic SMOTE in certain degree.

Hierarchical Multi-label Classification (HMC)

Improved TPR Ensemble

- The True Path Rule (TPR) consistency ensemble.



Negative propagation

Positive propagation

$$\begin{cases} d_i = 1 \Rightarrow d_{par(i)} = 1 \\ d_i = 0 \Rightarrow d_{child(i)} = 0 \end{cases}$$

$$\varphi_i(x) = \{ j \mid j \in child(i), d_j(x) = 1 \}$$

$$ph_i(x) = w \cdot \overline{ph}_i(x) + \frac{1-w}{|\varphi_i(x)|} \sum_{j \in \varphi_i(x)} ph_j(x)$$

Drawback: in practice, the performance of binary SVM classifiers for class tree nodes are obviously different. Error predictions of poor classifiers will give an obvious effect to other node classifiers .

Hierarchical Multi-label Classification (HMC)

Improved TPR Ensemble

- **Improved TPR** ensemble strategy.

To restrict the poor performance binary classifiers bring an error propagation effect to other classifiers. Another weight of classifier performance is introduced to the positive prediction propagation.

$$ph_i(x) = w \cdot \overline{ph}_i(x) + \frac{1-w}{|\varphi_i(x)|} \sum_{j \in \varphi_i(x)} v_j \, ph_j(x)$$

$$v_j = \frac{E_j}{\sum_{k \in \varphi_i(x)} E_k}, \; j \in \varphi_i(x)$$

Weight of classifier performance

Hierarchical Multi-label Classification (HMC)

Experiments

- Benchmark FunCat yeast Datasets.

  Downloaded from: *http://dtai.cs.kuleuven.be/clus/hmcdatasets/*

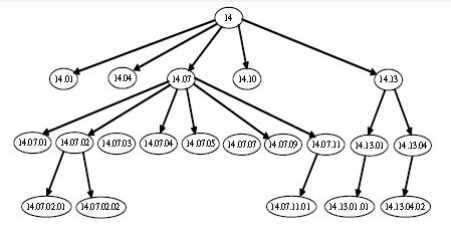
PROPERTIES OF EXPERIMENT DATASETS

Dateset	Attribute	Training	Testing
Sequence (seq)	478	2580	1339
Spellman et al. (cellcycle)	77	2476	1281
Gasch et al. (gasch1)	173	2480	1284
All microarray (expr)	551	2488	1291

Hierarchical Multi-label Classification (HMC)

Experiments

- Specific Experiments on Subtree Dataset
  - To observe the detailed performance of the proposed hierarchical SMOTE and the improved TPR ensemble, we also implement a specific experiment based on the dataset of Sequence (seq) and the subtree of “protein fate” FunCat class (FunCat ID = 14).
  - The subtree is composed by 20 nodes. In the dataset of Sequence (seq), there are 625 training examples and 337 testing instances relating with the class subtree of ID = 14.



Hierarchical Multi-label Classification (HMC)

Experiments

- Evaluation metrics

Per-class:

$$Prec = \frac{TP}{TP + FP}, Rec = \frac{TP}{TP + FN}$$

$$F - score = \frac{2 * Prec * Rec}{Prec + Rec}$$

Multi-label:

$$MPrec = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|y_i \cap z_i|}{|z_i|}, MRec = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|y_i \cap z_i|}{|y_i|}$$

$$MF - score = \frac{2 * MPrec * MRec}{MPrec + MRec}$$

Hierarchical Multi-label Classification (HMC)

Experiments

- Experiment results of different over-sampling strategies.

Table 4.2: Experiment results of different over-sampling strategies based on the FunCat subtree of "Protein fate" (ID=14).

FunCat ID	Positive	Negative	Flat Prec	Flat Rec	SMOTE Prec	SMOTE Rec	H-SMOTE Prec	H-SMOTE Rec
14.01	61	564	0.1667	0.3333	0.1721	0.3784	0.2407	0.4415
14.04	158	467	0.3131	0.3690	0.3371	0.3814	0.3558	0.4405
14.07	306	319	0.6460	0.5474	0.6460	0.5474	0.6460	0.5474
14.07.01	14	292	0.1020	0.4167	0.2141	0.4072	0.2111	0.4833
14.07.02	41	265	0.0822	0.2500	0.0978	0.2571	0.1486	0.4583
14.07.02.01	11	30	0.0714	0.7500	0.1521	0.4860	0.0882	0.7500
14.07.02.02	27	14	0.0606	0.8571	0.0606	0.8571	0.0606	0.8571
14.07.03	91	215	0.2619	0.6226	0.2987	0.6233	0.2736	0.6604
14.07.04	32	274	0.1019	0.5500	0.0925	0.3672	0.1400	0.7000
14.07.05	44	262	0.0532	0.2174	0.0872	0.2556	0.0860	0.3478
14.07.07	11	295	0.0121	0.1453	0.0195	0.1670	0.0263	0.3333
14.07.09	6	300	0.0153	0.1475	0.0205	0.1892	0.0385	0.2500
14.07.11	49	257	0.0690	0.2857	0.0944	0.3034	0.0909	0.5000
14.07.11.01	13	36	0.0244	0.2500	0.0215	0.1931	0.0357	0.3750
14.10	118	507	0.2474	0.4068	0.2850	0.4768	0.2700	0.4576
14.13	150	475	0.3048	0.4507	0.2715	0.5272	0.3380	0.5070
14.13.01	112	38	0.1688	0.7407	0.1275	0.4248	0.1688	0.7407
14.13.01.01	75	37	0.1147	0.7143	0.1147	0.7143	0.1147	0.7143
14.13.04	16	134	0.0333	0.1429	0.0314	0.1762	0.0556	0.2857
14.13.04.02	6	16	0.0127	0.2542	0.0127	0.2542	0.0127	0.2542

Hierarchical Multi-label Classification (HMC)

Experiments

- Experiment results of different ensemble strategies.

Table 4.3: Experiment results of different ensemble strategies based on the FunCat subtree of "Protein fate" (ID=14).

FunCat ID	TPR Prec	TPR Rec	Improved TPR Prec	Improved TPR Rec
14.01	0.2016	0.3792	0.2578	0.4356
14.04	0.3228	0.3929	0.3521	0.4568
14.07	0.5490	0.8263	0.6705	0.8947
14.07.01	0.2000	0.4167	0.2176	0.5116
14.07.02	0.0923	0.2500	0.2275	0.4835
14.07.02.01	0.1667	0.5000	0.2483	0.5758
14.07.02.02	0.0794	0.3571	0.1829	0.7683
14.07.03	0.2857	0.6038	0.2837	0.6487
14.07.04	0.0824	0.3500	0.1285	0.6982
14.07.05	0.0685	0.2174	0.0920	0.4276
14.07.07	0.0121	0.1453	0.0303	0.3333
14.07.09	0.0153	0.1475	0.0657	0.3150
14.07.11	0.0749	0.2786	0.1267	0.546
14.07.11.01	0.0169	0.1436	0.0465	0.3542
14.10	0.2474	0.4068	0.3043	0.4792
14.13	0.2516	0.5634	0.3357	0.5547
14.13.01	0.1825	0.4630	0.1864	0.6932
14.13.01.01	0.1221	0.7429	0.1237	0.7265
14.13.04	0.0378	0.1865	0.0567	0.2487
14.13.04.02	0.0127	0.2542	0.0476	0.5000

Hierarchical Multi-label Classification (HMC)

Experiments

- Three multi-label metrics on the FunCat subtree dataset.

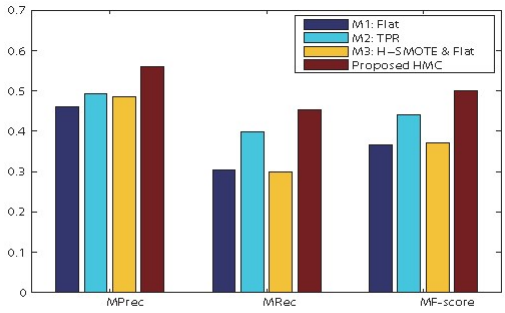


Figure 4.4: Three multi-label metrics on the FunCat subtree of ID=14 (Protein fate).

Hierarchical Multi-label Classification (HMC)

Experiments

- Classification results on the four FunCat datasets.

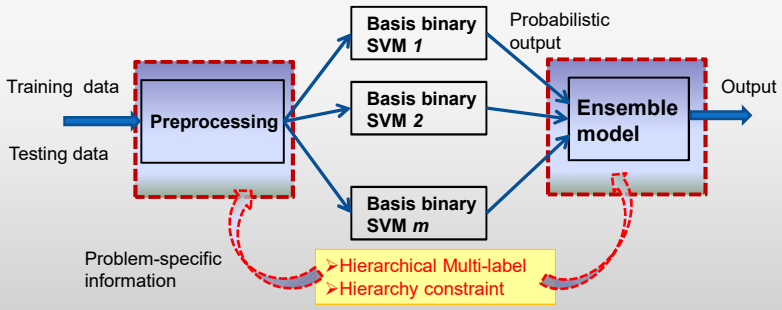
Table 4.4: Classification results on the four FunCat datasets.

Dataset	Method	MPrec	MRec	MF-score
D1	Flat	0.4771	0.3049	0.3720
	TPR	0.6056	0.3568	0.4490
	Proposed	0.6832	0.4362	0.5324
	Proposed + ck-SVM	0.6905	0.4487	0.5439
D2	Flat	0.4598	0.2876	0.3539
	TPR	0.5529	0.3624	0.4378
	Proposed	0.6513	0.4292	0.5174
	Proposed + ck-SVM	0.6597	0.4472	0.5331
D3	Flat	0.4694	0.2917	0.3598
	TPR	0.5883	0.3342	0.4263
	Proposed	0.6728	0.4318	0.5260
	Proposed + ck-SVM	0.6704	0.4342	0.5270
D4	Flat	0.4628	0.2932	0.3590
	TPR	0.5785	0.3427	0.4304
	Proposed	0.6593	0.4186	0.5121
	Proposed + ck-SVM	0.6685	0.4276	0.5216

Hierarchical Multi-label Classification (HMC)

Summary

1) A **hierarchical SMOTE** approach is proposed to preprocess the imbalanced training subsets for classifiers.
2) An **improved TPR consistency approach** is used to combine the results of binary probabilistic SVM classifiers. A **performance weight** is introduced into TPR method to restrict the error propagation effect s of poor performance binary classifiers.



Conclusions

1) A multi-label classification based on label ranking and delicate decision boundary SVM can be used for solving multi-label gene function classification.
2) A hierarchical multi-label classification (HMC) method based on over-sampling and hierarchy constraint can be used for solving the FunCat gene function prediction problem.

Thank you for your attention !

- End -