САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
Кафедра «Прикладная математика»

РАЗРАБОТКА АЛГОРИТМА АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ ИЕРАРХИЙ ИЗ НАБОРА ДАННЫХ ДЛЯ КЛАССИФИКАЦИИ ДОКУМЕНТОВ

Диссертация на соискание ученой степени магистра

Выполнил студент гр. 6057/2 Г.А.Сапожников
Научный руководитель А.В.Уланов

2012 г.

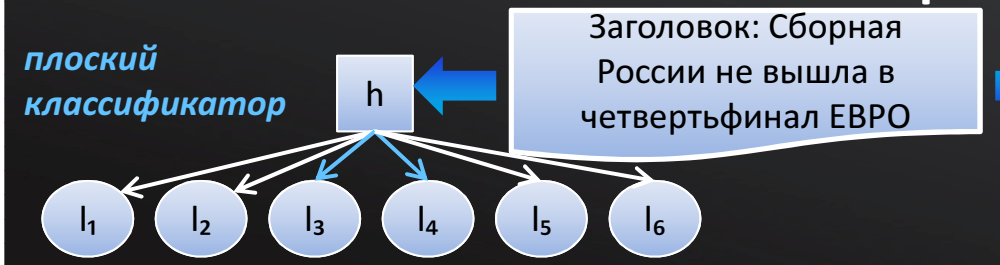
План доклада

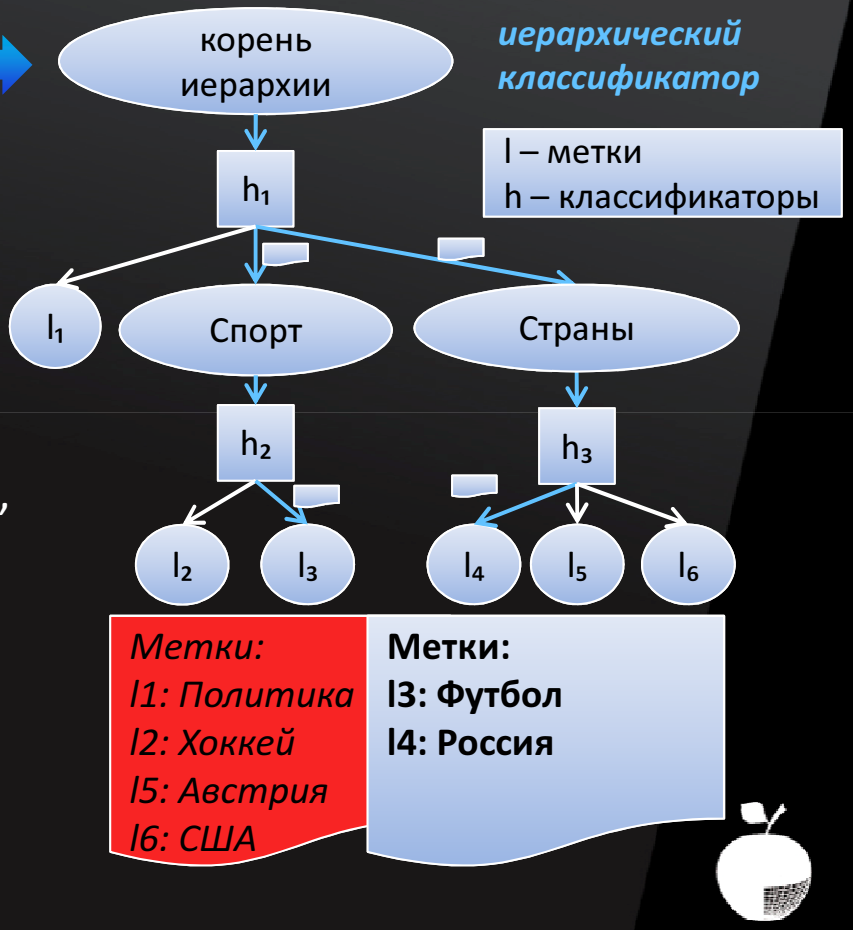
– Введение
– Многозначная классификация
– Мотивация
– Постановка задачи
– Решение задачи, алгоритм
– Результаты экспериментов
– Заключение

Введение

- Наборы данных
- Информационный поиск
- Классификация
- Иерархическая классификация

Многозначная классификация

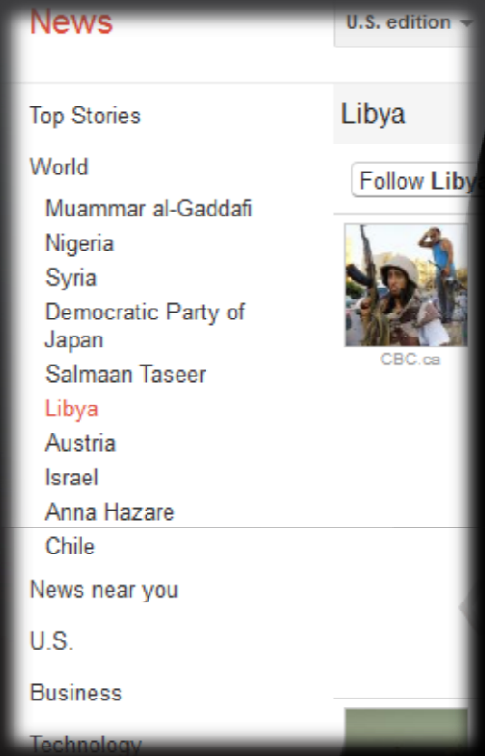




- Если каждый документ в задаче классификации может иметь более одной метки, это – многозначная классификация
- Метки могут иметь иерархическую структуры, которую можно использовать для классификации
- В иерархии документ попадает в корень. В каждом узле классификатор принимает решение, в какие узлы документ пойдет далее

Мотивация

- Почему иерархическая классификация может быть лучше:
  - Лучший баланс обучающих примеров
  - Каждый классификатор в иерархии имеет дело с меньшим количеством меток, по сравнению с полным набором
- Иногда наборы данных уже имеют иерархию (WIPO, Wikipedia, RCV, Google news)
- Что делать, если иерархии нет?
  - Алгоритм HOMER, ECML/PKDD'08 [Тсумакас и др., 08]
  - Угадывание иерархии [Брюкер и др., 11]
  - Иерархия из существующих меток [Ванг и др., 11]



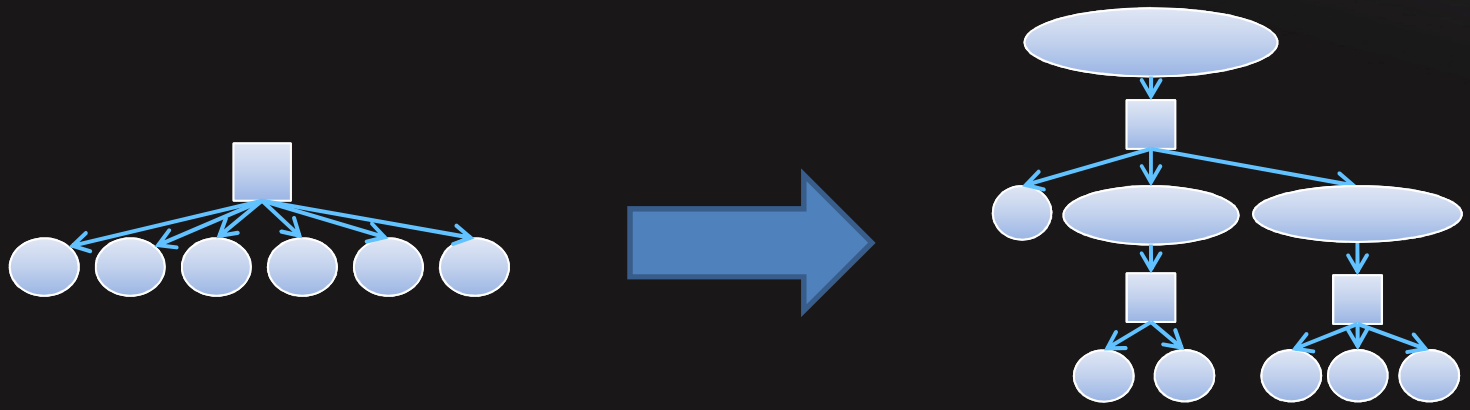
Bloomberg

WIPO World Intellectual Property Organization

LexisNexis

Постановка задачи

- **Задача**: Построение иерархии меток из набора данных. В дальнейшем иерархия будет использована для классификации документов
- **Дано:** Размеченные наборы документов
- **Идея:** Трансформация задачи многозначной классификации с большим количеством меток в иерархию более простых задач многозначной классификации



Используемый подход

Предлагается интеллектуальный алгоритм для построения иерархий для классификации:

– Автоматическое построение иерархии для классификации с помощью кластеризации

– Использование критериев, оптимизирующих различные меры: точность (P), полноту (R) или F-меру (F1)

– Оптимизируется иерархия, а не алгоритм классификаторов

TP – верно проставленные метки
FP – ошибочно проставленные метки
FN – ошибочно не проставленные метки

$$P = \frac{TP}{TP + FP}; R = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R}$$

Алгоритм: построение иерархии по слоям

Иерархия строится сверху вниз

1. Имеется некоторое количество меток (классов)
2. Производится различная кластеризация меток (различное количество кластеров)
3. Предсказывается, какое из разбиений будет лучшим для классификации (следующий слайд)
4. Процесс продолжается рекурсивно для кластеров, содержащих более двух меток

Алгоритм: функция предсказания

Задача: оценить, какое из разбиений будет лучшим для классификации, используя информацию о текущей точности классификации и оценивая точность классификации на последующих слоях

Пример: оценить F1 меру для разбиения l_1 l_2, l_3 l_4, l_5, l_6

1. Известны размеры всех классов и кластеров
2. Известны результаты классификации по кластерам на данном уровне
3. Делается оценка точности классификации внутри каждого из кластеров
   - В результате классификации мы знаем все о l_1
   - Для l_2, l_3 и l_4, l_5, l_6 вычисляются оценки TP, FN и FP
4. Можно вычислить оценку TP, FP, FN для всего разбиения
5. Вычисляется F1 мера для всего разбиения

Критерий остановки: оценка точности классификации всех кластеров ниже результата плоской классификации на текущем уровне

Результаты экспериментов

- Оптимизация F1 меры
- F1 была оптимизирована для четырех наборов
- В узлах иерархии использовались деревья решений C4.5[Webb.99]
- Все наборы документов взяты с http://mulan.sourceforge.net/
  - Текст: Bibtex, Enron, Medical
  - Музыка: CAL500
  - Изображения: Corel5k, Scene
  - Биология: Genbase, Yeast
  - Видео: Mediamill

Имя набора данных	Кол-во меток	Плоский			Иерархия		
		F1	P	R	F1	P	R
Mediamill	101	0,54	0,66	0,45	0,53	0,55	0,52
Bibtex	159	0,31	0,81	0,19	**0,39**	0,65	0,28
Medical	45	0,80	0,85	0,75	**0,83**	0,86	0,79
Enron	53	0,46	0,66	0,35	**0,50**	0,55	0,46
Yeast	14	0,59	0,61	0,58	0,59	0,59	0,60
Genbase	27	0,97	1,00	0,95	*0,98*	1,00	0,96
Corel5k	374	0,04	0,24	0,02	**0,09**	0,18	0,06
CAL500	174	0,37	0,5	0,3	*0,40*	0,37	0,43
Scene	6	0,61	0,67	0,57	*0,63*	0,66	0,59

Выводы

- Результаты:
  - Был предложен новый алгоритм для автоматического построения иерархий для классификации, оптимизирующий ее эффективность
  - Алгоритм основан на подборе наилучшего разбиения кластеров для любого уровня иерархии
  - Исследование алгоритма на девяти наборах документов показало его эффективность по сравнению с плоской классификацией. Точность классификации в среднем выросла на 3%
  - Результаты опубликованы в трудах конференций DEXA'11 и SDM'12
- Дальнейшая работа:
  - Использование алгоритма для масштабных наборов документов
  - Использование различных классификаторов в иерархии
  - Перестроение неудачных частей у наборов документов с иерархий

Спасибо за внимание!

Дополнительные слайды

Алгоритм: сравнение с HOMER

- При некоторых параметрах разрабатываемый алгоритм становится алгоритмом HOMER
- Если построить множество иерархий алгоритмом HOMER можно получить схожие результаты с исследуемым алгоритмом на некоторых иерархиях, однако не существует метода выбора между всеми иерархиями HOMER до использования тестового множества

Пример сравнения:

	HOMER									Алг.
	2	3	4	5	6	7	8	9	10	
Bibtex	0,35	0,38	0,38	0,37	0,38	0,38	0,38	**0,39**	0,38	**0,39**
Mediamill	0,49	0,51	0,52	**0,53**	**0,53**	**0,53**	**0,53**	**0,53**	**0,53**	**0,53**

Используемые меры

TP – верно проставленные метки
FP – ошибочно проставленные метки
FN – ошибочно не проставленные метки
K – количество классов
индекс ^{i} – значения для i-го класса

$$P^i = \frac{TP^i}{TP^i + FP^i}; R^i = \frac{TP^i}{TP^i + FN^i}$$

$$F1^i = \frac{2 \cdot P^i \cdot R^i}{P^i + R^i}$$

$$P_{micro} = \frac{\sum_{i=1}^{K} TP^i}{\sum_{i=1}^{K} TP^i + \sum_{i=1}^{K} FP^i}; R_{micro} = \frac{\sum_{i=1}^{K} TP^i}{\sum_{i=1}^{K} TP^i + \sum_{i=1}^{K} FN^i}$$

$$F1_{micro} = \frac{2 \cdot P_{micro} \cdot R_{micro}}{P_{micro} + R_{micro}}$$

$$P_{macro} = \frac{\sum_{i=1}^{K} P^i}{K}; R_{macro} = \frac{\sum_{i=1}^{K} R^i}{K}$$

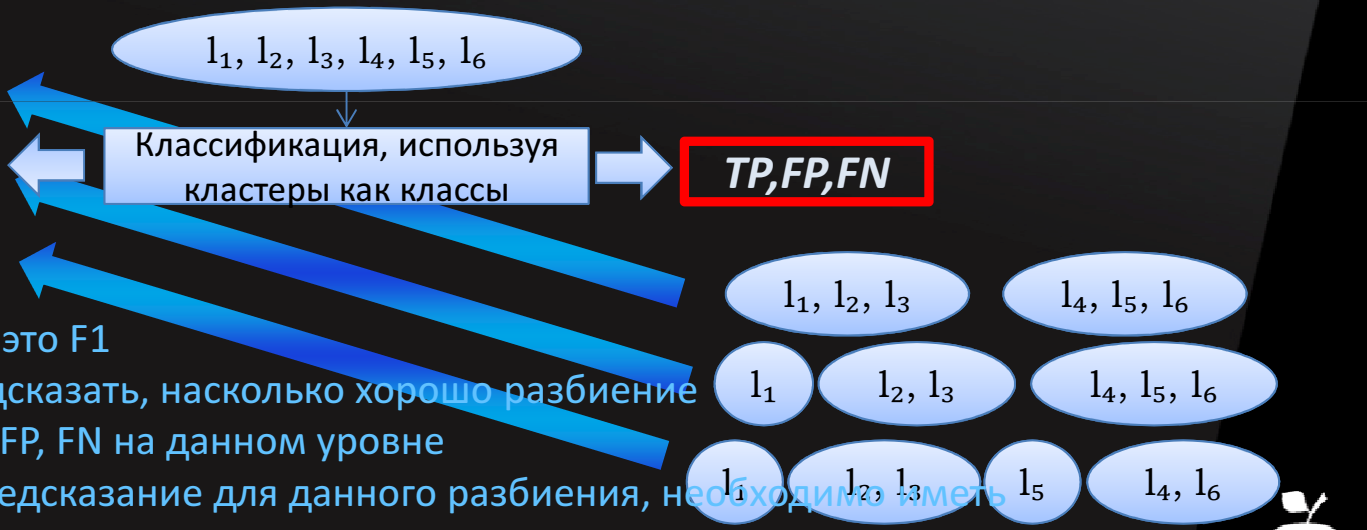
$$F1_{macro} = \frac{\sum_{i=1}^{K} F1^i}{K}$$

Функция предсказания (шаг 1)

Необходимо предсказать, какое из разбиений будет наилучшим для классификации, принимая во внимание **точность классификации по кластерам на данном уровне** и дальнейшую достройку иерархии

Данный уровень: Классификация документов с использованием кластеров в качестве мета-меток. В результате получаем значения мер точности классификации по кластерам

C	P^T	R^T
2	0,99	0,95
3	0,96	0,90
4	0,92	0,88



Пример: оптимизации это F1

- Необходимо предсказать, насколько хорошо разбиение
- Известны P, R, TP, FP, FN на данном уровне
- Чтобы сделать предсказание для данного разбиения, необходимо иметь ввиду не только результаты на данном уровне, но и последующие уровни