О классификации текстов в условиях неполноты обучающего множества

А.Ю. Колесов

Ярославский государственный университет им. П.Г. Демидова

20 октября 2011 г.

Задача классификации

Рассматриваем задачу тематической классификации текстовых документов.

- Классификация/рубрикация – отнесение объектов к одной или нескольким рубрикам из конечного множества рубрик.
- Дано: множество категорий C и обучающее множество – пары релевантности документ-рубрика.
- Требуется приблизить неизвестную функцию

$$\Phi : D \times C \to \{0, 1\},$$

используя обучающее множество. D – множество текстовых документов.

Шаги решения задачи классификации

Напомним эти шаги (для индуктивного подхода).

1. Преобразование: документ $\longrightarrow$ математический объект (вектор в n-мерном пространстве).
2. Восстановление функции Φ по обучающему множеству.
3. Используя восстановленную функцию $\hat{\Phi}$, классифицируем новые объекты.
4. Оцениваем качество классификации.

Актуальность

1. Низкое качество классификации для мультиклассовых задач (менее 50% полноты и точности), несмотря на большое количество методов.
   - Большое количество объектов – часто большое количество рубрик.
   - Пересечения рубрик: один объект может принадлежать ко многим классам.
2. Дорогостоящий процесс создания обучающего множества для таких задач.
   - Для документа необходимо среди большого числа рубрик отметить полный набор рубрик, релевантных документу. (Иначе документ, принадлежащий классу A, но не отмеченный меткой этого класса, попадает в отрицательные примеры для класса A).
   - Не всегда это условие выполняется и является приемлемым.

Предположения

1. Существуют объекты из обучающего множества, составленного экспертами, имеющие неполный набор меток, и их достаточно много.
2. Эксперты не ошибаются: любая метка, которую эксперт проставил для документа, правильна.
3. Гипотеза компактности: «Схожие объекты гораздо чаще лежат в одном классе, чем в разных; или, другими словами, что классы образуют компактно локализованные подмножества в пространстве объектов».

Предлагаемый подход

Задача модификации обучающего множества: для каждой рубрики требуется найти те документы из обучения, которые, исходя из геометрии данных, релевантны этой рубрике, и добавить их в обучение.
Для решения необходимо:

1. С помощью специального алгоритма получаем возможные пары релевантности документ-рубрика (множество AD), которые не отметили эксперты.
2. Используем один из способов применения множества AD:
   - Добавляем множество AD в обучение.
   - Для пар $(d, c) \in \text{AD}$ – при обучении для рубрики c документ d не попадет в отрицательные примеры для этой рубрики.

Новая схема решения задачи

1. Преобразование: документ $\longrightarrow$ математический объект (вектор в n-мерном пространстве).
2. Модификация обучающего множества согласно предлагаемому методу.
3. Восстановление функции Φ по модифицированному обучающему множеству.
4. Используя восстановленную функцию $\hat{\Phi}$, классифицируем новые объекты.
5. Оцениваем качество классификации.

Алгоритмы модификации: Soft-supervised learning (I)

Предполагается, что все объекты, которые требуется отклассифицировать, известны. На вход алгоритму подается множество из размеченных и неразмеченых объектов $D = \{D_l, D_u\}$. Каждому такому объекту сопоставляется набор вероятностей принадлежности классам $p_i = (p_i^t)_{t=1}^m$, где m – количество классов.

Алгоритмы модификации: Soft-supervised learning (II)

Задача сводится к минимизации функционала $C_1(p)$ по наборам вероятностей $p = (p_1, \ldots, p_n)$:

$$\min_p C_1(p), \text{ где } C_1(p) = \sum_{i=1}^{l} D_{KL}(r_i||p_i) + \\ + \mu \sum_{i=1}^{n} \sum_{j \in K(i)} w_{ij} D_{KL}(p_i||p_j) - \nu \sum_{i=1}^{n} H(p_i)$$

$D_{KL}(p_i||p_j)$ – расстояние Kullback-Leibler, $H(p_i)$ – энтропия. $r_i = (r_i^t)_{t=1}^m$ – известный набор вероятностей (то, что разметили эксперты).
Вес w_{ij} определяется так: $w_{ij} = \mathrm{sim}(d_i, d_j)\delta(j \in K(i))$, где $K(i)$ – множество k ближайших соседей объекта d_i.

Алгоритмы модификации: Soft-supervised learning (III)

После минимизации функционала $C_1(p)$ мы получаем для каждого документа $d_i \in D$ набор вероятностей $p_i = (p_i^1, \ldots, p_i^m)$. Далее вводится порог $T \in [0, 1]$, с помощью которого выделяем дополнительные рубрики, релевантные документу: документ $d_i \in c_j$, если $p_i^j \geq T$.

Алгоритмы модификации: Weighted KNN

$$p_i^j = \frac{\sum\limits_{t \in \mathrm{K}(i)} \mathrm{sim}(d_i, d_t)\varphi(d_t, c_j)}{\sum\limits_{t \in \mathrm{K}(i)} \mathrm{sim}(d_i, d_t)},$$

где $\varphi(d_t, c) = 0$, если $d_t \notin c$, и $\varphi(d_t, c) = 1$, если $d_t \in c$.
Также вводим порог $T \in [0, 1]$: документ $d_i \in c_j$, если $p_i^j \geq T$.

Эксперименты. Данные

Опишем эксперименты на коллекции:

- Agingportfolio.
  - База данных проектов, связанных со старением и финансируемых Национальным институтом здоровья (NIH) и Европейской комиссией (EC CORDIS).
  - Более 1 млн. 100 тыс. записей о научных проектах (название, краткое описание, теги).
  - Средняя длина документа – 100 слов.
  - 335 рубрик на 6 уровнях иерарархии.

Эксперименты. Обучение и тест Agingportfolio

Тестовое множество:

- Для каждого документа из теста имелось два набора категорий, составленных разными экспертами.
- В качестве меток взято объединение этих наборов. Это позволяет получить более полный набор категорий.
- Тестовое множество включало 750 проектов.

Обучающее множество:

- Обучающее множество составлено с меньшим контролем, разными людьми, в том числе пользователями ресурса Agingportfolio. Как показывает визуальный анализ, в обучающем множестве довольно большое количество проектов имеет неполный набор категорий.
- Среднее число рубрик на проект в обучающем множестве составляет 4,36, а на тестовой – 9,79.
- Обучающее множество включало 3144 проектов.

Алгоритм обучения

Алгоритм обучения – линейный SVM.

- Подход one-vs-rest для реализации мультиклассовой классификации.
- Решающее правило: $w_c x + b_c > 0$ для каждого класса c.
- Применялись:
  - SVM с параметрами по умолчанию.
  - SVM с подбором параметров.

Подбор параметров SVM

Подбирались следующие параметры:

- параметр C (характеризует компромисс между максимизацией разделяющей полосы и минимизацией суммарной ошибки);
- параметр b (порог b_c классификации в решающем правиле).

Подбор параметров осуществлялся методом скользящего контроля с разбиением множества на 5 частей (5-fold cross-validation).

Результаты и выводы

Таблица: **Результаты на Agingportfolio**

	macro_f	micro_f
noFit-SVM	2,02%	22,64%
SVM	14,61%	28,47%
del+w-kNN	16,76%	24,34%
add+w-kNN	20,15%	39,43%
del+SoftSL	15,09%	20,37%
add+SoftSL	16,11%	31,13%

Таблица: **Среднее количество рубрик на документ и документов на рубрику в различных обучающих множествах. Коллекция Agingportfolio**

	avg_rubr_cnt	avg_doc_cnt
no_add_modif	4,36	44,35
add+w-kNN	11,86	120,62
add+SoftSL	10,78	111,05

Таблица: **Полнота и точность результатов классификации. Коллекция Agingportfolio**

	micro_recall	micro_prec
noFit-SVM	13,00%	87,77%
SVM	22,86%	37,73%
del+w-kNN	50,08%	16,07%
add+w-kNN	39,39%	39,46%
del+SoftSL	39,65%	25,62%
add+SoftSL	46,90%	13,01%

Заключение

- Предложен подход для повышения эффективности применения алгоритма SVM в задаче классификации в условиях неполного набора меток объектов из обучения.
- Эффективность подхода показана на задаче классификации научных грантов с большим числом рубрик.
- Метод на основе w-kNN и стратегии «добавление релевантных документов в обучающее множество» дает улучшение качества классификации по сравнению с базовым (SVM) по F_1-мере на 38% и при макроусреднении, и при микроусреднении.

Конец

- Вопросы?

Список литературы

1. Esuli A, Sebastiani F: **Training Data Cleaning for Text Classification**. In *Proceedings of the 2nd International Conference on the Theory of Information Retrieval (ICTIR'09)*, Cambridge, UK 2009:29–41.
2. J Tang AF Z Chen, Cheung D: **Capabilities of outlier detection formulation schemes, framework and methodologies**. *Knowledge and Information Systems* 2007, **11**:45–84. [Springer].
3. N G Zagoruiko VVD I A Borisova, Kutnenko OA: **Methods of recognition based on the function of rival similarity**. *Pattern Recognition and Image Analysis* 2008, **18**:1–6.
4. Lam Hong Lee TFY Chin Heng Wan, Kok HM: **A Review of Nearest Neighbor-Support Vector Machines Hybrid Classification Models**. *Journal of Applied Sciences* 2010, **17**:1841–1858.
5. Subramanya A, Bilmes J: **Soft-Supervised Learning for Text Classification.** In *EMNLP'08* 2008:1090–1099.
6. Subramanya A, Bilmes JA: **Entropic Graph Regularization in Non-Parametric Semi-Supervised Classification**. In *Neural Information Processing Society (NIPS)*, Vancouver, Canada 2009.

Список литературы-2

7. Manning CD, Raghavan P, Schütze H: *An Introduction to Information Retrieval*, Cambridge, England: Cambridge University Press 2009 .
8. **International aging research portfolio**. *http://agingportfolio.org* . [Accessed 25 April 2011].
9. Robertson SE, Walker S, Jones S, Hancock-Beaulieu M, Gatford M: **Okapi at TREC-3**. In *Proceedings of the Third Text REtrieval Conference (TREC 1994)* 1994.
10. Joachims T: **A Statistical Learning Model of Text Classification with Support Vector Machines**. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)* 2001:128–136.
11. Schölkopf B, Smola AJ, Williamson RC, Bartlett PL: **New Support Vector Algorithms**. *Neural Computation* 2000, **12**:1207–1245.
12. **Chih-Jen Lin's Home Page.** *http://www.csie.ntu.edu.tw/~cjlin/index.html* . [Accessed 28 July 2011].
13. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ: **LIBLINEAR: A Library for Large Linear Classification**. *Journal of Machine Learning Research* 2008, **9**:1871–1874.