Обработка и классификация документов с использованием системы СКАТ

© Васильев В.Г.

ООО «ЛАН-ПРОЕКТ»
vvg_2000@mail.ru

Аннотация

В статье рассматриваются методы классификации и процедуры обработки обучающего множества текстов, которые реализованы в системе СКАТ. Приводятся итоги обработки документов в рамках дорожки классификации нормативно-правовых документов.

Введение

Целью участия в семинаре РОМИП-2009 являлась оценка эффективности средств классификации текстов, реализованных в системе СКАТ на стандартном тестовом массиве большого объема. При проведении экспериментов ставились следующие задачи: отладка процедур обучения классификаторов на больших объемах документов; оценка эффективности использования различных процедур обработки обучающего множества текстов; оценка эффективности совместного использования нескольких методов классификации; оценка степени совпадения показателей качества классификации текстов, получаемых в результате самооценки с использованием метода кросс-проверки и получаемых в результате использования официального тестового множества.

Работа имеет следующую структуру. В первом разделе приводятся краткое описание общей технологии классификации текстов в системе СКАТ и реализованные алгоритмы обработки обучающего множества текстов. Во втором разделе приводятся результаты оценки качества классификации.

1. Описание методов классификации

1.1 Общее описание метода классификации

В системе СКАТ используется метод комбинированной иерархической классификации [2], который основан на одновременном использовании нескольких методов классификации и объединении результатов их работы. Такой подход позволяет автоматически осуществлять выбор наиболее подходящего метода для каждой рубрики и учитывать наличие зависимостей между рубриками (например, учитывать иерархическую структуру дерева рубрик).

Обработка текстов осуществляется с использованием трех уровней классификаторов:

1. базовых классификаторов рубрик;
2. комбинированных классификаторов рубрик;
3. общего иерархического классификатора.

Базовые классификаторы рубрик [1]. На первом уровне для каждой рубрики осуществляется построение набора бинарных классификаторов (осуществляют разбиение документов на два класса) с использованием следующих базовых методов классификации:

- методов вероятностной классификации, основанных на представлении рубрик в виде смеси распределений Бернулли (BERN), фон Мизеса-Фишера (VMF), полиномиального (MNS) и анализаторов главных компонент (PPCA);
- методов классификации на основе вычисления расстояний: классификаторы k-ближайших соседей (KNN), машин опорных векторов (SVM), Роччио (ROC);
- методов классификации на основе правил: деревья решений (TREE);
- методов классификации на основе запросов: построение запросов на специальном информационно-поисковом языке.

Для возможности комбинирования результатов работы данных методов в системе произведена унификация технологии их обучения и интерфейса взаимодействия.

Комбинированные классификаторы рубрик [1]. На втором уровне для каждой рубрики осуществляется построение комбинированного бинарного классификатора. Для этих целей поддерживается использование нескольких типов методов: с фиксированным решающим правилом, основанных на оценках качества, основанных на использовании статистического

моделирования. При этом основным является метод, основанный на выборе наилучшего метода классификации для каждой рубрики на основе оценок качества классификации.

Для оценки качества классификации документов поддерживается использование нескольких методов: самооценка, перепроверка, к-шаговая кросс-проверка. Последний метод используется в качестве основного метода в экспериментах, проводимых в настоящей работе.

Общий иерархический классификатор. Обеспечивает интеграцию результатов работы комбинированных классификаторов отдельных рубрик и общее управление процессом обучения и классификации документов. В частности, на данном уровне осуществляется оценка адекватности входных текстов обучающему множеству, формирование обучающих и тестовых множеств для классификаторов нижнего уровня, корректировка результатов работы классификаторов отдельных рубрик.

1.2 Описание методов обработки обучающего множества

Результаты предварительных экспериментов по обучению классификатора нормативно-правовых документов ROMIP-2007 выявили ряд особенностей его структуры и содержания, которые осложняют прямое использование стандартных методов классификации текстов. В частности, были выявлены следующие особенности обучающего множества:

- имеются ошибки в распределении документов по рубрикам;
- имеется большое количество рубрик, которые дублируют друг друга (например, состав обучающих примеров полностью совпадает для следующих рубрик: c848 и c874; c10 и c267; c9 и c102; c1033, c1552, c790, c947; c1034, c791, c858, c836 и др.);
- имеется большое количество рубрик с очень похожими названиями и содержанием, распределение документов по которым является неоднозначным;
- большое количество рубрик (порядка 700) и количество текстов (несколько десятков тысяч), в этой ситуации количество отрицательных и положительных примеров при обучении отдельных рубрик является непропорциональным, а время обучения классификаторов сильно возрастает;

Большинство распространенных обучаемых методов классификации (SVM, деревья решений и др.) являются достаточно чувствительными к наличию указанных особенностей и ошибок в обучающем множестве текстов. По этой причине в настоящей работе с целью повышения итогового качества классификации

текстов было решено оценить эффективность использования различных процедур коррекции обучающего множества текстов.

Необходимо отметить, что в системе СКАТ при обучении базового классификатора для отдельной рубрики все множество обучаемых примеров разбивается на два класса относящихся (положительные примеры) и не относящихся (отрицательные примеры) к данной рубрике. При этом множество отрицательных примеров, как правило, оказывается в несколько десятков раз больше множества положительных примеров.

По этой причине было решено рассмотреть различные процедуры отбора только отрицательных примеров и оценить их влияние на качество классификации. Для реализации различных процедур отбора отрицательных примеров в системе СКАТ были реализованы следующие базовые операции.

Таблица 1. Операции отбора отрицательных примеров

Операция	**Описание**
$\text{negative}(S)$	Отбор из множества S всех отрицательных текстов (не являются положительными для заданной рубрики).
$\text{common}(S,a)$	Отбор из множества S всех текстов, которые содержат не более a процентов терминов из словаря терминов, которые встречаются в положительных примерах.
$\text{dsim}(S,a)$	Отбор из множества S всех текстов, для которых не существует положительных примеров, косинусное расстояние до которых меньше заданного порога.
$\text{rsim}(S,a)$	Отбор из множества S всех текстов, которые не относятся к рубрикам, косинусное расстояние от центра которых до центра заданной рубрики меньше a.
$\text{random}(S,a,b)$	Отбор из множества S случайного подмножества, содержащего не более, чем b текстов и не более, в чем a раз больше текстов, чем во множестве положительных примеров.
$\text{sib}(S)$	Отбор из множества S всех текстов, которые находятся в смежных рубриках для заданной рубрики (данная операция применяется для иерархических

	классификаторов).
$\text{near}(S,a)$	Отбор из множества S подмножества a текстов, имеющих наибольшее число терминов из словаря терминов положительного множества текстов.
$\text{far}(S,b)$	Отбор из множества S подмножества b текстов, имеющих наименьшее число терминов из словаря терминов положительного множества текстов.

Выбор указанных операций для отбора отрицательных примеров был обусловлен следующими соображениями.

Операция $\text{negative}(S)$ соответствует варианту, когда используются все отрицательные примеры. Использование данной операции соответствует традиционному случаю.

Операции $\text{dsim}(S,a)$, $\text{rsim}(S,a)$, $\text{common}(S,a)$ реализуют отбрасывание отрицательных примеров, которые слишком похожи на положительные примеры. Такие примеры очень часто соответствуют текстам, которые ошибочно не отнесены к положительным примерам или являются дубликатами положительных примеров. Первоначально были реализованы первые две операции. Однако как показали эксперименты, при обработке большого массива текстов операция $\text{dsim}(S,a)$ требует $O(n^2)$ операций вычисления расстояний, что является достаточно обременительным. По это причине была реализована значительно более эффективная операция $\text{common}(S,a)$, основанная на оценке доли новых слов в тексте по отношению к словарю терминов для заданной рубрики.

Операция $\text{random}(S,a,b)$ обеспечивает сокращение размера множества отрицательных примеров путем случайного выбора подмножества текстов заданного размера. Данная операция необходима в тех случаях, когда размер множества отрицательных примеров становится очень большим и время обучения классификатора для рубрики значительно увеличивается.

Операции $\text{sib}(S)$, $\text{near}(S,a)$, $\text{far}(S,b)$ также предназначены для сокращения размера множества отрицательных примеров, однако за счет отбора документов, удовлетворяющих определенным свойствам. В частности, операция $\text{sib}(S)$ в качестве примеров

отбирает документы из соседних рубрик, которые потенциально расположены более близко к множеству положительных примеров. Операции $\mathrm{near}(S,a)$ и $\mathrm{far}(S,b)$ отбирают наиболее близкие и далекие примеры соответственно. Однако в вычислительном плане они являются ресурсоемкими, так как требуют вычисления расстояний между элементами множеств положительных и отрицательных примеров.

Эксперименты

1.1 Описание предварительных экспериментов

Для построения процедуры обработки обучающего множества текстов было проведено экспериментальное исследование различных вариантов ее построения. Описание оцениваемых процедур и параметров отдельных операций приведено в следующей таблице.

В данной таблице используются следующие обозначения:

C – обучающее множество с плоской структурой рубрик, для которого выполнена автоматическая очистка документов от элементов оформления и заголовков (данная процедура по умолчанию применяется для очистки страниц из Интернет от второстепенной информации);

T – обучающее множество с плоской структурой рубрик, для которого не выполнялась очистка документов;

H – обучающее множество с двухуровневой структурой рубрик, для которого не выполнялась очистка документов;

R – итоговое множество отобранных отрицательных примеров.

Таблица 2. Описание процедур обработки обучающего множества

Название	**Описание**
scat_1	R=Random(Negative(C), 5, 1000)
scat_2	C1 = Common(Negative(C), 99%) R=Random(C1, 20, 2000)
scat_3	C1 = Common(Negative(C), 99%) R=Random(H1, 20, 7000)
scat_4	C1 = Common(Negative(C), 95%) R=Random(C1, 20, 7000)
scat_5	H1 = Common(Negative(H), 95%) R=Random(H1, 20, 7000)

hier_scat_6	T1 = Sib(Negative(T)) T2 = Common(T1, 95%) R=Random(T2, 20, 7000)

Для самооценки качества классификации использовался метод 5-шаговой кросс проверки. Результаты проведенных экспериментов приведены в следующей таблице.

Таблица 3. Результаты самооценки качества классификации на полном множестве рубрик

Эксперимент	**F**	**Precision**	**Recall**
scat_1	11%	6%	58%
scat_2	17%	11%	49%
scat_3	29%	27%	31%
scat_4	25%	24%	27%
scat_5	44%	35%	57%
hier_scat_6	45%	45%	45%

По результатам указанных экспериментов можно сделать следующие выводы:

- выполнение отбрасывания ближайших отрицательных примеров с использованием процедуры Common позволяет значительно улучшить качество классификации;
- выполнение дополнительной очистки HTML документов только ухудшает качество классификации (это, скорее всего, связано с тем, что в данном случае в документах нет избыточной и шумовой информации);
- учет иерархической структуры рубрик позволяет незначительно улучшить качество классификации (это может быть связано с тем, что в данном случае рубрики имеют не иерархическую, а фасетную организацию).

1.2 Описание результатов официальных экспериментов

С учетом результатов, полученных при предварительных экспериментах, на официальную оценку было решено отправить результаты классификации массива текстов, полученные с использованием классификатора, построенного в рамках эксперимента «scat_5».

В следующей таблице приводятся соответствующие результаты официальных экспериментов, полученные по дорожке тематической классификации нормативно-правовых документов.

Таблица 4. Общие результаты тематической классификации нормативно-правовых документов

Характеристика	**XXXX-1**	**scat_5**
F1 (micro average)	15.5%	23.5%
Recall (micro average)	9.9%	45.1%
Precision (micro average)	35.2%	15.9%
F1	16.9%	20.4%
Recall	14.5%	48.5%
Precision	33.5%	16.3%
Accuracy	98.7%	99.2%
Error	1.3%	0.8%
Not judged	5638	67420

В целом анализ данной таблицы позволяет сделать вывод, что качество классификации в обоих случаях оказывается достаточно низким. При этом у системы СКАТ средние показатели качества оказываются несколько более высокими, чем у первой системы.

При проведении предварительных экспериментов вычисление итоговых показателей качества проводилось на полном множестве из 720 рубрик, а при официальной оценке только на случайном подмножестве из 75 рубрик. Для возможности проведения сравнения официальных и предварительных результатов в следующей таблице приводятся соответствующие средние значения показателей качества для 75 рубрик.

Таблица 5. Средние значение показателей точности и полноты классификации для 75 рубрик

	Precision	**Recall**
Самооценка	0,36	0,57
Официальная оценка	0,16	0,49

Оценки, приведенные в таблице, показывают, что результаты, получаемые при самооценке, являются завышенными. При этом более сильно искажено значение точности классификации. Возможной причиной такого поведения оценок может быть плохое соответствие распределения текстов по классам в обучающем и тестовом множестве.

Заключение

Эксперименты, проведенные в работе, показали, что выполнение дополнительных процедур по очистке обучающего множества текстов позволяет значительно повысить итоговое качество классификации.

Важными моментами при практической реализации средств классификации также являются следующие: компактность представления текстов в памяти компьютера (при достаточно большом объеме обучающего множества данные могут не помещаться в оперативную память); состав признаков, которые используются для представления текстов, и выполняемые процедуры по предварительной обработке самих текстов; вычислительная сложность алгоритма классификации (обработка тестового массива занимает достаточно продолжительное время).

К перспективным направлениям дальнейших исследований и экспериментов можно отнести следующие:

- проведение анализа эффективности использования итерационных методов обработки обучающего множества текстов, которые зависят от оценок качества классификации;
- проведение анализа объемно-временных характеристик алгоритмов построения классификаторов;
- оценка эффективности использования обратной связи с пользователями (оценщиками) при обучении классификатора, апробация различных методов активного обучения;
- оценка эффективности учета не только иерархических, но фасетных взаимосвязей между рубриками.

Работа выполнена при поддержке гранта Президента Российской Федерации для государственной поддержки молодых российских ученых МК-12.2008.10.

Литература

[1] Васильев, В.Г., Кривенко, М.П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008. – 302 с.

[2] Васильев, В.Г. Комплексная технология автоматической классификации текстов // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции "Диалог". – М. РГГУ, Вып. 7(14), 2008. - с. 83-90.

Document processing and classification in SCAT system

Vitaly Vasilyev

The paper describes the learnable methods of texts classification and procedures of training set preprocessing which are realized in SCAT system. Results of legal track processing are discussed.