Оценка применимости генетических алгоритмов в целях сужения пространства признаков рубрик в задачах автоматической классификации текстовых данных

© Дивинский А. П., Бабичев Н. В.

«Группа Вито»
divinskiy@gmail.com, sharky@pcn.ru

Аннотация

В данной работе авторы попытались оценить возможности и перспективы применения генетических алгоритмов в целях сужения пространства признаков рубрик для задач автоматической классификации текстовых данных. Оценка результатов эксперимента проводилась в рамках семинара РОМИП`2006.

1. Введение

Если рассматривать задачу сужения пространства признаков документа как комбинаторную задачу с несколькими оптимумами, то генетические алгоритмы представляются довольно эффективным средством ее решения, поскольку являются методами стохастического, эвристического поиска. В то же время ГА обладают определенными недостатками, такими как достаточно высокая вычислительная сложность и чувствительность вызывающая тенденцию к переобучению при использовании их в качестве средства реализации машинного обучения.

На данный момент нет математически обоснованной теории строго формулирующей критерии которым должна соответствовать задача, для того чтобы можно было с уверенностью сказать, что она решается эффективно с помощью генетических алгоритмов [1]. Однако, анализ результатов практического использования ГА

позволяет выделить условия, при которых имеет смысл их применение [2]:

- большое пространство поиска которое содержит несколько экстремумов;
- многокритериальность поиска;
- поиск приемлемого решения по заданным критериям в отличие от поиска единственного оптимального.

Исходя из этого, ГА представляются приемлемым методом решения поставленной задачи.

2. Коллекция нормативно-правовых документов

Прежде чем перейти к описанию принципов работы системы и результатов прогонов, стоит сказать несколько слов о самой коллекции нормативно-правовых документов и обучающем множестве для нее.

Для 60208 документов, которые требуется классифицировать, определено 183 рубрики (класса). Один документ может входить одновременно в несколько классов – от 1-го до 5-и.

В качестве обучающего множества предлагается коллекция из 6293 документов, для которых указана принадлежность к классам рубрикатора. Таким образом, косвенным описанием каждого класса является набор документов, представляющий собой подмножество документов обучающей коллекции.

Количество документов-примеров	Количество рубрик
0	10
1	11
2	10
3	8
4	6
5	5
6-10	19
11-20	21
21-50	31
>50	62

Таблица 1. Характеристика обучающего множества.

Обучающее множество для коллекции нормативно-правовых документов, своей неоднородностью представляет определенные сложности для обучения классификатора (*таблица 1*) (см. также [3]).

В данном эксперименте было решено рассматривать рубрики, содержащие более четырех документов в своем описании. Их количество составило 133. Максимальное количество документов-примеров в рубрике было ограничено до 100.

3. Краткое описание метода

3.1 Предварительная обработка документов

Предварительная обработка документов включает следующие операции:

- удаление форматирования документа, цифровых и специализированных символов;
- приведение всех символов документа к нижнему регистру;
- определение идентификаторов лемм;
- построение индекса.

3.2 Начальное формирование множеств ключевых слов рубрик

Исходя из того факта, что словарный запас и частоты использования слов зависят от тематики документа, для каждой рубрики R_i множество ключевых термов $K(R_i)$ строится на основе статистического анализа встречаемости термов в документах обучающего множества $D(C)$ [4].

Чем выше частота встречаемости документов $\wp(t, R_i)$ содержащих терм t на множестве документов $D(R_i)$ входящих в рассматриваемую рубрику, относительно частоты встречаемости документов содержащих терм t на множестве $D(C) \setminus D(R_i)$, тем выше вес $w_{R_i}(t)$ рассматриваемого терма для рубрики R_i .

Определим частоту терма t на множестве документов D как

$$\wp(t, D) = \frac{\left|\{d \mid d \in D \wedge t \in d\}\right|}{|D|} \qquad (1)$$

Как видим, $\wp(t, D)$ представляет собой вероятность появления термина t в случайно выбранном документе из множества D.

Пусть

$$\wp_{R_i} = \wp(t, D(R_i))$$

$$\wp_{C_i} = \wp(t, D(C) \setminus D(R_i))$$

Введем эвристическую формулу для определения веса терма в рубрике:

$$w_{R_i}(t) := \left\{ \begin{array}{c} \log_a \left(\frac{\wp_{R_i}}{\wp_{C_i}} \right), \wp_{C_i} > \frac{\{n \mid n > 0\}}{|D(C)|} \\ 0, \wp_{C_i} \le \frac{\{n \mid n > 0\}}{|D(C)|} \end{array} \right\} \qquad (2)$$

Решение использовать в формуле (2) $\wp_{C_i}$ вместо $\wp_C = \wp(t, D(C))$ обусловлено желанием предотвратить влияние неоднородности обучающего множества на вес терма относительно рубрики в тех случаях, когда количество документов представляющих класс R_i в обучающем множестве составляет существенную часть от общего количества документов в коллекции.

Для формирования множества ключевых термов $K(R_i)$ рубрики выбираем термы такие, что

$w_{R_i}(t) > b, b > 0$ (позитивные признаки)

или

$w_{R_i}(t) < c, c < 0$ (негативные признаки).

В описываемом эксперименте мы использовали значения *a=e, b=1.5*. В данном эксперименте негативные признаки было решено не использовать, так как результаты предварительных экспериментов показали недостаточно высокую эффективность их применения.

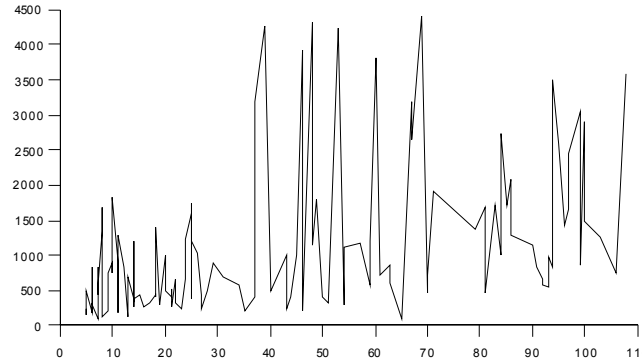

График 1. Зависимость количества термов выбранных в качестве ключевых для рубрик от количества документов-примеров характеризующих рубрики в обучающем множестве

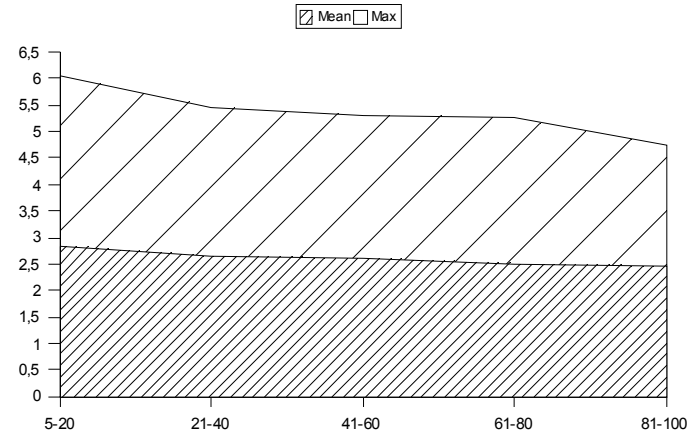

График 2. Зависимость среднего и максимального значений весов термов выбранных в качестве ключевых для рубрик от количества документов-примеров характеризующих рубрики в обучающем множестве

Как показано на *графике 1*, при использовании формулы (2), стабильной зависимости значения $|K(R_i)|$ от количества документов-примеров не наблюдается. Среднее значение $avg(K(R_i))$ не зависит от $D(R_i)$, а максимальное значение веса терма в рубрике с увеличением количества документов-примеров падает незначительно (*график 2*).

3.3 Классификация документов

Коэффициент релевантности рубрики для документа определяется формулой

$$k_i(R_i) = \sum_{\forall t \in K(R_i)} w_{R_i}(t) w_d(t)$$

где $w_d(t)$ - вес терма в документе.

Релевантными для документа ***d*** считаются такие рубрики, что

$$R_i \in \left\{ R \mid \frac{k_i(R_i)}{max(K)} > \varepsilon \right\} \qquad (3)$$

3.4 Применение ГА для сужения пространства признаков рубрики

Реализация ГА для рассматриваемой задачи тривиальна и выходит за рамки этой статьи.

Опишем основной момент – выбор оценочной функции качества множества признаков рубрики $S(R_i)$.

Допустим S – это набор множеств признаков $S(R)$ для всех рубрик коллекции. Задача ГА состоит в сокращении размерности каждого множества $S(R_i)$ до $m \le |S(R_i)| \le M$ стремясь при этом достигнуть оптимального значения оценочной функции для качества полученного множества.

Вполне очевидным решением представляется в роли оценочной функции качества множества $S_x(R_i)$ использовать значение $F1$ для результатов классификации множества документов D используя

классификатор S, в котором текущая рубрика R_i представлена рассматриваемым множеством $S_x(R_i)$.

Для рассматриваемой задачи двумя самыми очевидными способами задания оценочной функции качества множества $S_x(R_i)$ являются:

$$G_1(K_x(R_i)) = FI(T(D(R_i))) \quad (4)$$

или

$$G_2(K_x(R_i)) = FI(T(D(C))) \quad (5)$$

где $T(D)$ представляет собой таблицу результатов классификации [5] множества документов D классификатором S в котором рубрика R_i представлена множеством признаков $S_x(R_i)$.

Ясно, что время вычисления функции G_1 меньше времени вычисления G_2, причем при работе ГА для каждой рубрики в среднем

$$\frac{time(G_1)}{time(G_2)} \sim \frac{|D(R_i)|}{|D(C)|}$$

В тоже время, использование функции G_1 предположительно даст худшее качество полученных в итоге наборов признаков документа, чем использование функции G_2, по той причине, что, при определении качества множества признаков каждой рубрики, функция G_1 не учитывает возможность попадания в рассматриваемый класс, документов из обучающего множества относящихся к другим рубрикам, что неблагоприятно сказывается на точности результатов классификации.

В нашем случае признаками рубрик являются их ключевые термы выбранные методом описанным в *параграфе 3.2*.

4. Анализ результатов

4.1 Описания прогонов

На семинаре РОМИП`2006 эксперимент был представлен результатами трех прогонов:

1. NF – прогон классификатора с использованием в качестве признаков рубрик множества термов отобранных способом описанным в параграфе 3.2
2. GAN – прогон классификатора с использованием в качестве признаков рубрик суженные с помощью ГА до мощности 40 множества термов использовавшихся в прогоне NF . В качестве оценочной функции для работы ГА использовалась функция G_1
3. AF – был осуществлен с целью проверить предположение о том, что вероятность попадания документа в рубрику зависит не только от величины коэффициента релевантности определенного методом описанным в данной статье, но и от количества документов которыми рубрика представлена в обучающем множестве. Это предположение было сделано на основании того, что в рассматриваемой задаче один документ может принадлежать нескольким рубрикам, и, возможно, рубрика, которая содержит большее количество документов, содержит меньшее количество документов относящихся только к ней, таким образом являясь предположительно общей для некоторых других рубрик коллекции. Формула (2) была модифицирована соответствующим образом.

4.2. Сравнение результатов

Как видно на графике 3, для прогона NF не наблюдается устойчивой корреляции между значениями метрик и количеством документов-примеров в обучающей коллекции. Из этого можно сделать вывод, что функция (2) чувствительна не столько к количеству документов представляющих рубрику, сколько к структуре самого обучающего множества: насколько документы представляющие ту или иную рубрику типичны для нее, какое количество рубрик в коллекции терминологически близки друг другу. Таких результатов можно было ожидать исходя из анализа *графиков 1 и 2.*

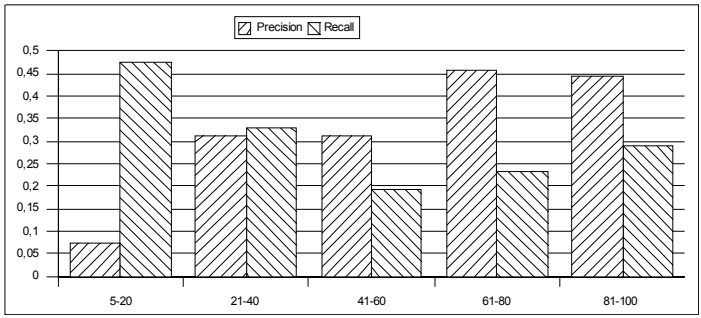

График 3. Диаграмма зависимости значений метрик от количества обучающих примеров для прогона NF

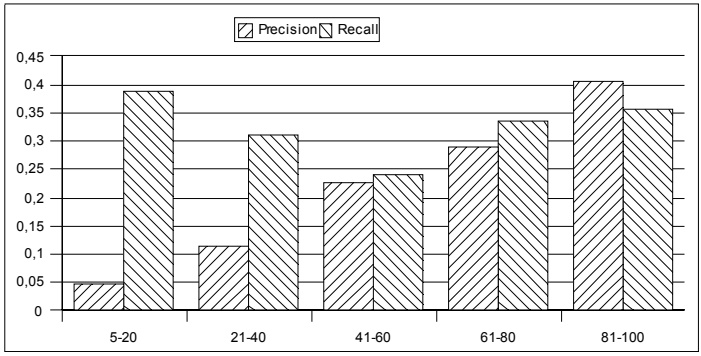

График 4. Диаграмма зависимости значений метрик от количества обучающих примеров для прогона GAN

В результатах прогона GAN по сравнению с результатами прогона NF значение полноты снизилось незначительно. Авторы считают это вполне удовлетворительным результатом при учете того, что мощность множеств ключевых термов документов снизилась в среднем в 26.81 раз. В тоже время, значение точности уменьшилось по сравнению с результатами погона NF более существенно, что обусловлено недостатками функции G_1 (см. *параграф 3.4*). Авторы предполагают, что при использовании в

качестве оценочной функции G_2, значение точности должно оказаться близко к значению полноты.

Результаты прогона AF показали более высокую относительно прогонов NF и GAN точность – около 0.28, а полнота оказалась несколько ниже и также приближенно равна 0.28. Подробный анализ результатов этого прогона выходит за рамки этой статьи.

4.3 Рассмотрение результатов прогона GAN относительно результатов прогонов других систем

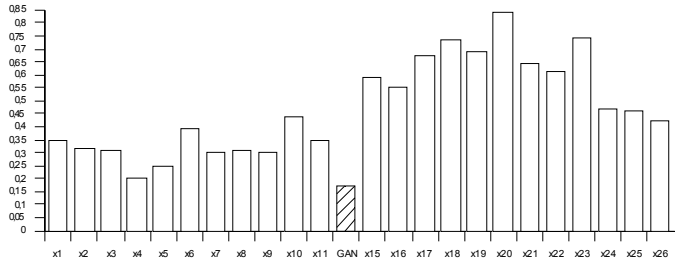

График 5. Диаграмма точности результатов прогонов

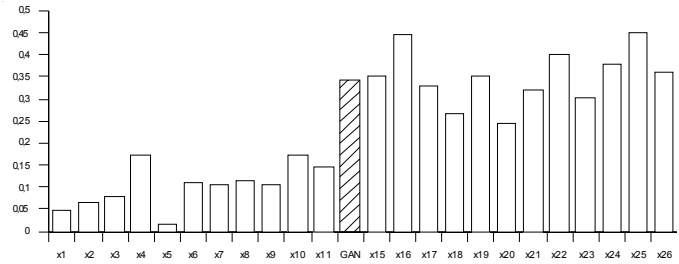

График 6. Диаграмма полноты результатов прогонов

Как показывают *графики 5 и 6* результат прогона GAN по точности является наиболее низким из всех результатов достигнутых системами участвовавшими в семинаре РОМИП`2006. Причины этого рассмотрены выше. В тоже время, результат прогона GAN входит в число лучших по полноте.

5. Выводы и дальнейшие планы

По мнению авторов, результаты оценки работы описанного метода полученные на семинаре РОМИП`2006 показали перспективность применения генетических алгоритмов в целях сужения пространства признаков рубрик в задачах автоматической классификации текстовых данных.

Слабые показатели точности в результатах прогона GAN обусловлены, в основном, недостатками функции G_1 и негибкостью (3).

Дальнейшие эксперименты с ГА планируется провести в следующих направлениях:

- применение G_2 в качестве оценочной функции для ГА;
- применение ГА для коррекции весов признаков рубрик;
- разработка комбинированных методов.

Также, планируется разработка более эффективных методов выделения признаков рубрик.

Литература

[1] *Джордж Ф. Люгер*. Искусственный интеллект. Стратегии и методы решения сложных проблем (Artificial Intelligence. Structures and Strategies for Complex Problem Solving) – Вильямс, 2003 – 864 стр.

[2] *Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И.* Методы и модели анализа данных: OLAP и Data Mining – БХВ-Петербург, 2004 – стр. 299-303

[3] *Агеев М. С., Добров Б. В., Лукшевич Н. В., Сидоров А. В.* Экспериментальные алгоритмы поиска/классификации и сравнение с «basic line». // Российский семинар по Оценке Методов Информационного Поиска (РОМИП 2004) – Пущино, 2004.

[4] *Некрестьянов И. С.* Тематико-ориентированные методы информационного поиска: Дис. канд. физ-мат. Наук: 05.13.11 / С-Пб. гос. унив. – Санкт-Петербург, 2002

[5] *И. Кураленок, И. Некрестьянов* Оценка систем текстового поиска. / Программирование. – 28(4), 2002

The applicability estimation of genetic algorithms for space narrowing of rubrics` signs in tasks of automatic classification of text data

© A. Divinsky, N. Babichev

"Vito`s Group"

divinskiy@gmail.com, sharky@pcn.ru

The authors tried to estimate abilities and perspectives of genetic algorithms application with the aims of narrowing rubrics` signs for tasks of automatic classification of text data. The estimation of the experiment results was carrying out within the bounds of seminar RIRES`2006.