электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.852

Определение полезности признаков в задаче классификации коротких текстовых сообщений

Дремина А.К., *студент*
кафедра «Информационные системы и телекоммуникации»,
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана

Научный руководитель: Павлов Ю. Н., д.т.н., профессор
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
deviatkov@bmstu.ru

Методология оценки полезности признаков. Обоснование подхода

Пусть X_n – обучающая выборка, в которой все объекты описываются признаками $(x_1, x_2, \dots, x_n)$, а X_{n-1} – та же самая выборка, но без последнего признака. Пусть m – некоторый метод обучения по прецедентам. То есть $a = m(X_n) : X \to Y$ – алгоритм классификации. Пусть также имеется некоторый функционал качества классификации $F(a)$. То есть классификация тем лучше, чем большее значение принимает $F(a)$. Полезностью фактора x_n будем называть величину $\mathrm{p} = \mathrm{F}(\mathrm{m}(\mathrm{X_n})) - \mathrm{F}\big(\mathrm{m}(\mathrm{X_{n-1}})\big)$. Нетрудно заметить, что эта величина означает приращение функционала качества при добавлении фактора в обучающую выборку.

Отметим также, что эта величина зависит не только от самого признака, но также и от того, какие признаки уже присутствуют в выборке. Поэтому для исследования полезности различных нетривиальных признаков, добавим в обучающую выборку признак, численно равный результату применения классического текстового классификатора – алгоритма наивного Байеса. При таком подходе, полезность остальных признаков, поочередно добавляемых в выборку, будет показывать, насколько увеличивается качество классификации благодаря данному признаку в сравнении с классическим методом. Для количественного измерения полезности признака, нам остается выбрать функционал качества классификации и метод обучения.

Функционал качества классификации

Точность и полнота являются классическими метриками, которые используются для оценки качества классификаторов. Точность классификатора в пределах класса – это доля документов действительно принадлежащих данному классу относительно всех

документов, которые классификатор отнес к этому классу. Полнота – это доля найденных классификатором документов, принадлежащих классу, относительно всех документов этого класса. Удобно определить следующие величины:

- TP — истинно-положительное решение;
- TN — истинно-отрицательное решение;
- FP — ложноположительное решение;
- FN — ложноотрицательное решение.

Тогда, точность и полнота определяются следующим образом:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Понятно, что чем выше точность и полнота, тем лучше. Однако зачастую приходится искать баланс между этими величинами. Поэтому удобно ввести величину, которая бы объединяла в себе информацию как о точности классификатора, так и о его полноте. Именно такой метрикой является F-мера. Она представляет собой гармоническое среднее между точностью и полнотой. Заметим, что она стремится к нулю, если точность или полнота стремится к нулю.

$$F = \frac{2(Precision \cdot Recall)}{Precision + Recall}$$

В данной работе качество классификации измеряется именно F-мерой. В качестве метода обучения используется классический алгоритм классификации – метод опорных векторов.

Пунктуационные факторы сообщений и эмотиконы

Особенностью коротких электронных сообщений является использование эмотиконов («смайликов») – символов, которые описывают эмоции человека. В данной работе в качестве пунктуационных признаков учитывались следующие эмотиконы:

- количество эмотиконов, обозначающих улыбку – **:)**
- количество эмотиконов, обозначающий подмигивание – **;)**
- количество эмотиконов, обозначающий грусть, уныние – **:(**
- количество эмотиконов, обозначающий смех – **:D**

Также особенностью коротких текстовых сообщений является отсутствие пунктуации (опускание точек, запятых), повторение знаков пунктуации (к примеру, повторение восклицательных знаков), опускание пробелов перед и после знаков препинания.

В качестве пунктуационных признаков использовались следующие факторы:

- количество точек;

- количество точек, после которых идет пробел;
- количество точек, до которых идет пробел;
- количество запятых;
- количество запятых, после которых идет пробел;
- количество запятых, до которых идет пробел;
- количество восклицательных знаков;
- количество вопросительных знаков;
- количество тройных восклицательных знаков («!!!»);
- количество тройных вопросительных знаков («???»).

Описание данных

В качестве данных используются сообщения пользователей сервиса Twitter (www.twitter.com). Данный сервис позволяет отправлять короткие текстовые сообщения сообщения (до 140 символов), используя веб-интерфейс, SMS, сторонние программы-клиенты.

Полученные результаты

Описанная схема была применена к оценке двухклассовой классификации авторов с помощью анализа пунктуационных факторов коротких текстов – твитов пользователей, полученных с помощью Twitter API. Результаты представлены в Таблице 1.

Полученная F-мера классификации с помощью наивного байесовского классификатора колеблется между 71 % и 92 %. Среднее значение F-меры – 81.1 %. При этом самыми различимыми по точности авторами оказались @max_katz (муниципальный депутат, урбанист) и @arttema (дизайнер). При рассмотрении Twitter timeline обоих авторов оказалось, что @arttema обычно публикует очень короткие сообщения (содержащие 15-20 символов), которые состоят из слов и ссылки и при этом не содержат пунктуационных знаков и эмотиконов, в то время, как у пользователя @max_katz сообщения обычно занимают 100-140 символов и содержат много пунктуационных знаков и эмотиконов. Самыми схожими авторами при использовании наивного байесовского классификатора оказались @navalny и @max_katz. Примечательным является тот факт, что оба автора придерживаются одних политических взглядов и активно пишут об этом в Twitter.

Полученная F-мера классификации с применением наивного байесовского классификатора и факторов пунктуации колеблется между 78% и 98%. Среднее значение F-меры – 89%. Самыми различимыми авторами, как и при наивной байесовской классификации, являются пользователи @max_katz и @arttema, а самыми схожими –

@navalny (оппозиционер, адвокат) и @segalovich (технический специалист). При этом качество классификации по этим двум авторам выросло с 73.10% до 78.27%. Оба автора активно используют классическую пунктуацию и ставят пробелы после запятых.

В среднем же качество классификации при добавлении факторов пунктуации улучшилось согласно F-мере на 7.9%.

Таблица 1

Полученные результатов двухклассовой классификации текстов

	Байесовский классификатор			Байесовский классификатор + пунктуация			Вес фактора
	Точност	Полнота	F-мера	Точност	Полнота	F-мера	
@elkasinger + @navalny	86,25%	87,39%	86,82%	88,77%	90,41%	89,58%	2,76%
@elkasinger + @M_Galustyan	76,05%	76,05%	76,05%	87,87%	87,88%	87,87%	11,82%
@elkasinger + @twicehti	79,35%	82,98%	81,12%	86,82%	88,30%	87,55%	6,43%
@elkasinger + @max_katz	80,83%	83,77%	82,28%	88,16%	88,33%	88,25%	5,97%
@elkasinger + @arttema	89,88%	89,86%	89,87%	95,01%	95,44%	95,22%	5,35%
@elkasinger + @MedvedevRussia	81,75%	96,22%	88,39%	89,68%	97,79%	93,56%	5,17%
@elkasinger + @segalovich	82,42%	83,38%	82,90%	87,70%	87,66%	87,68%	4,78%
@navalny + @M_Galustyan	84,12%	84,79%	84,45%	88,94%	89,90%	89,42%	4,97%
@navalny + @twicehti	65,22%	82,67%	72,91%	81,14%	82,72%	81,92%	9,01%
@navalny + @max_katz	63,54%	80,97%	71,21%	82,37%	83,15%	82,76%	11,55%
@navalny + @arttema	91,32%	89,51%	90,41%	92,72%	90,73%	91,71%	1,31%
@navalny + @MedvedevRussia	74,33%	90,96%	81,81%	84,89%	92,33%	88,46%	6,65%
@navalny + @segalovich	67,49%	79,74%	73,10%	75,58%	81,16%	78,27%	5,17%
@M_Galustyan + @twicehti	77,83%	82,27%	79,99%	89,89%	90,03%	89,96%	9,98%
@M_Galustyan + @max_katz	79,66%	83,04%	81,31%	90,05%	90,17%	90,11%	8,79%
@M_Galustyan + @arttema	86,00%	86,45%	86,23%	95,90%	96,32%	96,11%	9,88%
@M_Galustyan + @MedvedevRussia	78,32%	88,79%	83,23%	88,19%	93,87%	90,94%	7,72%
@M_Galustyan + @segalovich	77,76%	80,52%	79,12%	88,18%	88,25%	88,21%	9,09%
@twicehti + @max_katz	87,40%	87,77%	87,58%	90,59%	90,54%	90,56%	2,98%
@twicehti + @arttema	91,67%	91,60%	91,63%	97,05%	97,31%	97,18%	5,55%
@twicehti + @MedvedevRussia	61,90%	92,66%	74,22%	78,65%	90,80%	84,29%	10,07%
@twicehti + @segalovich	82,03%	83,85%	82,93%	86,24%	86,53%	86,38%	3,45%
@max_katz + @arttema	92,85%	92,24%	92,54%	97,51%	97,59%	97,55%	5,00%
@max_katz + @MedvedevRussia	63,49%	92,36%	75,25%	85,92%	93,78%	89,68%	14,43%
@max_katz + @segalovich	78,38%	81,45%	79,89%	84,98%	85,47%	85,23%	5,34%
@arttema + @MedvedevRussia	91,87%	83,29%	87,37%	86,99%	95,59%	91,09%	3,72%
@arttema + @segalovich	90,07%	89,59%	89,83%	95,14%	95,87%	95,51%	5,68%
@MedvedevRussia + @segalovich	63,10%	87,10%	73,18%	78,39%	89,37%	83,52%	10,34%
Среднее	77,38%	86,76%	81,10%	86,29%	92,02%	89,00%	7,90%

Выводы

В результате выполненной работы был разработан и реализован метод оценки полезности факторов коротких сообщений при их классификации – идентификации авторов сообщений. Данный метод был применен для оценки полезности пунктуационных факторов сообщений, таких как знаки препинания и эмотиконы. Результаты показали, что при двухклассовой классификации пунктуационные факторы дают значимый прирост к качеству классификации. Это доказывает гипотезу о том, что пунктуация является важной и отличительной чертой коротких электронных сообщений разных пользователей.

В дальнейших исследованиях стоит рассмотреть подробнее пунктуационные факторы (добавить использование других символов пунктуации и других эмотиконов), а также проанализировать дополнительные стилевые факторы сообщений (использование заглавных букв, сокращений и т.д.). Перспективными также кажутся частотные факторы сообщений, основанные на словах, парах и тройках символов, и смысловые факторы, т.е. о чем пишет автор сообщения.

Список литературы

1. Информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. [электронный ресурс]: виртуальная энциклопедия – Режим доступа: http://machinelearning.ru/
2. Наивного байесовский классификатор [электронный ресурс]: виртуальная энциклопедия – Режим доступа: http://en.wikipedia.org/wiki/Naive_Bayes_classifier
3. Мещеряков Р.В., Романов А.С. Идентификация авторства коротких текстов методами машинного обучения [электронный ресурс]: портал международной конференции по компьютерной лингвистике Диалог – Режим доступа: http://www.dialog-21.ru/digests/dialog2010/materials/html/62.htm
4. История Твиттера [электронный ресурс]: виртуальная энциклопедия – Режим доступа: http://ru.wikipedia.org/wiki/Twitter
5. API [электронный ресурс]: виртуальная энциклопедия – Режим доступа: http://en.wikipedia.org/wiki/Application_programming_interface
6. API Twitter [электронный ресурс]: https://dev.twitter.com/
7. Метод опорных векторов [электронный ресурс]: виртуальная энциклопедия – Режим доступа: http://en.wikipedia.org/wiki/Support_vector_machine