УДК 004.89

АБДУКТИВНЫЙ ЛОГИЧЕСКИЙ ВЫВОД ДЛЯ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТОВ НА ОСНОВЕ ДСМ-МЕТОДА

Котельников Е.В.

ФГБОУ ВПО «Вятский государственный гуманитарный университет», Киров, e-mail: kotelnikov.ev@gmail.com

В статье предлагается абдуктивный метод логического вывода, предназначенный для анализа тональности текстов на основе ДСМ-метода автоматического порождения гипотез. Абдукция представляет собой процесс объяснения некоторого наблюдения в рамках заданной теории. В традиционной процедуре абдукции ДСМ-метода возникают проблемы при обработке коллекций текстов, связанные с большим количеством порождаемых гипотез, отсутствием обработки шумов и высокой вероятностью возникновения ситуации переобучения. Предлагаемый метод позволяет решить указанные проблемы за счет вычисления степени объясняющей способности гипотез и степени значимости обучающих объектов, а также на основе процедуры перекрестной проверки. Результатом работы метода является ранжированный по степени объясняющей способности список гипотез. Эксперименты с применением текстовой коллекции отзывов о фильмах семинара РОМИП-2011 подтверждают эффективность разработанного метода.

Ключевые слова: ДСМ-метод, абдукция, анализ тональности текстов

ADBUCTIVE LOGICAL INFERENCE FOR TEXT SENTIMENT ANALYSIS BASED ON JSM-METHOD

Kotelnikov E.V.

Vyatka State Humanities University, Kirov, e-mail: kotelnikov.ev@gmail.com

The abductive logical inference method for text sentiment analysis based on JSM-method for automatic hypotheses generation is proposed in the article. Abduction is the process of explanation of some observation in the context of a given theory. In the traditional abduction procedure of the JSM-method there are some problems in processing collections of texts related to the large number of generated hypotheses, lack of noise processing and a high probability of overfitting situation. The proposed method makes it possible to solve these problems by calculating the degree of explanatory ability of hypotheses and the degree of importance of training texts, as well as on the basis of cross-validation procedure. The result of this method is the list of hypotheses, which are ranked by the degree of explanatory ability. The experiments in which we use the text collections of movie reviews from seminar ROMIP-2011 confirm the effectiveness of the developed method.

Keywords: JSM-method, abduction, text sentiment analysis

Абдукция, или абдуктивный логический вывод, предложена американским философом Чарльзом Сандерсом Пирсом (1839–1914) в середине 1860-х гг. [10, 11]. В настоящее время общепринятым [7] является понимание абдукции как процедуры объяснения некоторого наблюдения в рамках заданной теории. Формальная постановка задачи предложена в [8]: при наличии теории T и наблюдения G задача абдукции заключается в обнаружении объяснения Δ, такого, что:

1) из $T \cup \Delta$ следует G;

2) T и Δ непротиворечивы.

В такой постановке задачи процедура абдукции включает три этапа [5, p. 9]:

1) генерация множества гипотез, подходящих на роль объяснений;

2) оценка гипотез;

3) принятие гипотез в качестве объяснений.

В разных работах по абдуктивному выводу акцент, как правило, ставится на один из этих этапов. Например, в статье [12] прогнозируется развитие ситуаций на основе абдуктивной генерации посылок; в статье [2] оценивается информативность гипотез для анализа тональности текстов. В ДСМ-методе автоматического порождения гипотез абдукция понимается как процедура принятия гипотез, объясняющих исходные факты [4, с. 18]. Гипотезы при этом порождаются в процессе индуктивного вывода.

Применение ДСМ-метода для анализа тональности текстов выявило огромное количество индуктивно генерируемых гипотез: при обработке порядка 10^4 текстов, каждый из которых содержит несколько десятков слов, может порождаться порядка 10^4–10^5 гипотез [3]. Такое количество гипотез не поддается непосредственной обработке исследователем в традиционном абдуктивном выводе. Другие проблемы, возникающие при использовании ДСМ-метода для обработки текстов, заключаются в отсутствии обработки шумовых объектов и высокой вероятности ситуации переобучения, вследствие того что абдукция применяется для принятия индуктивно порожденных гипотез, которые должны обобщать обучающие данные, с использованием тех же самых данных.

В данной статье предлагается метод абдуктивного вывода, позволяющий решить указанные проблемы:

1) обрабатывать множество гипотез большой мощности;

2) обнаруживать шумовые тексты и исключать их из обучающих данных;

3) нивелировать эффект переобучения.

Метод абдуктивного вывода

Метод абдуктивного вывода состоит из восьми основных этапов и представлен на рис. 1. На вход метода поступает обучающее множество текстов **T**, для которого требуется сформировать объясняющие гипотезы. На *первом этапе* осуществляется разделение обучающего множества случайным образом на q непересекающихся блоков $\mathbf{T}_i$, в каждом из которых содержится примерно одинаковое количество текстов. Значение q обычно выбирается равным 5 или 10. Такое разделение необходимо для процедуры перекрестной проверки (q-fold cross-validation) – стандартного способа предотвращения ситуации переобучения при построении систем машинного обучения [9]. В этой процедуре после разделения обучающего множества выполняется q потоков: при выполнении i-го потока i-й блок является контрольным: $\mathbf{T}_{test} = \mathbf{T}_i$, остальные блоки – обучающими: $\mathbf{T}_{train} = \bigcup_{j \neq i} \mathbf{T}_j$. В каждом потоке система использует для обучения множество $\mathbf{T}_{train}$, а для тестирования – $\mathbf{T}_{test}$. Оценки тестирования усредняются по всем потокам, и данный результат считается итогом процедуры перекрестной проверки с минимизацией вероятности ситуации переобучения.

Таким образом, для реализации процедуры перекрестной проверки этапы со второго по седьмой метода абдуктивного вывода выполняются независимо в q потоках.

На *втором этапе* происходит индуктивное порождение гипотез на основе обучающего множества текстов $\mathbf{T}_{train}$ с использованием, например, метода, предложенного в работе [1]. В результате формируется множество гипотез **H**, которые являются кандидатами в закономерности предметной области.

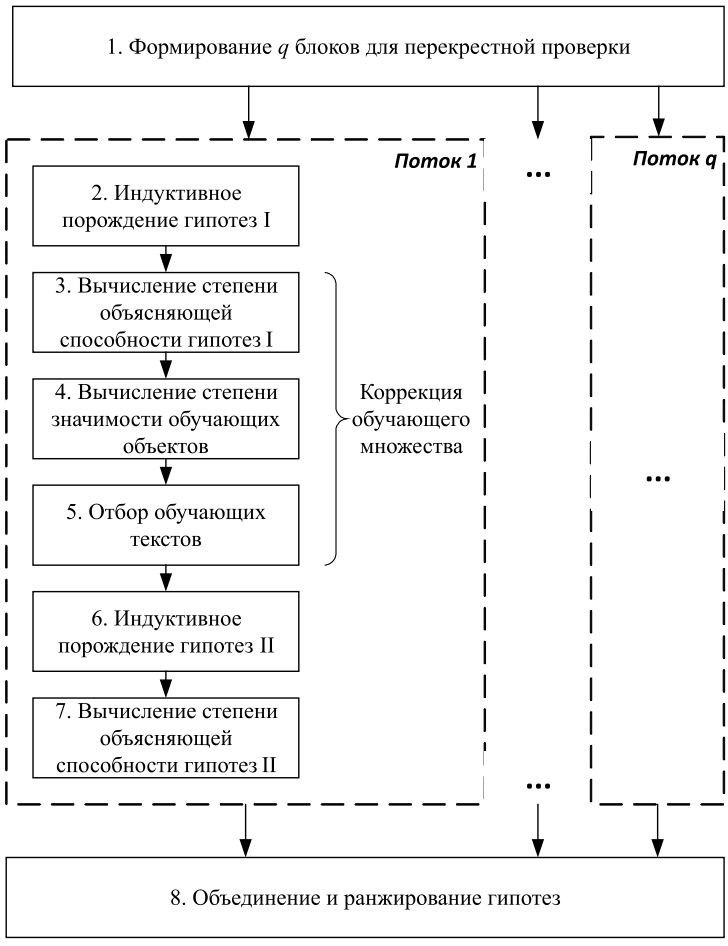

Рис. 1. Метод абдуктивного вывода

Этапы с третьего по пятый служат для коррекции обучающего множества. На *третьем этапе* для каждой гипотезы $h \in \mathbf{H}$ вычисляется степень её объясняющей способности *Abd* (от слова «abduction») для текстов контрольного множества $\mathbf{T}_{test}$ по следующей формуле:

$$Abd(h) = \sum_{t \in \mathbf{T}_{test}} SAW(h,t), \qquad (1)$$

где t – текст из контрольного множества $\mathbf{T}_{test}$, $SAW(h, t)$ – функция оценки информативности *SAW* (Sentiment analysis weight) гипотезы h относительно текста t, определяемая по формуле [2]

$$SAW = k_{sent} \cdot \frac{\log_2\left(2 + \frac{p}{\max(1,n)}\right)}{\log_2\left(Dist_{av} + 1\right)}, \qquad (2)$$

где k_{sent} – коэффициент оценочной лексики, учитывающий наличие в гипотезе h слов из словаря оценочной лексики; p – количество положительных текстов, распознаваемых гипотезой; n – количество отрицательных текстов, распознаваемых гипотезой; $Dist_{av}$ – среднее расстояние между словами гипотезы h в текущем тексте t.

Степени объясняющей способности всех гипотез, вычисленные по формуле (1), определяют применимость гипотез для описания обучающих текстов и позволяют выявить шумовые тексты на следующих этапах.

На *четвертом этапе* вычисляется степень значимости *Imp* (от слова «importance») каждого обучающего текста $t \in \mathbf{T}_{train}$ на основе степеней объясняющей способности порожденных данным текстом гипотез

$$Imp(t) = \frac{\sum_{h \in \mathbf{H}_t} Abd(h)}{Len(t)}, \qquad (3)$$

где $\mathbf{H}_t$ – множество гипотез, для генерации которых использовался текст t; $Len(t)$ – длина в словах текста t.

Вычисление степеней значимости по формуле (3) позволяет выявить обучающие тексты с низкими значениями $Imp(t)$, порождающие гипотезы с невысокими значениями $Abd(h)$, вследствие чего такие тексты можно считать шумовыми.

На *пятом этапе* тексты, признанные шумовыми в соответствии с заранее заданным порогом степени значимости, исключаются из обучающего множества $\mathbf{T}_{train}$. На *шестом этапе* повторно генерируются гипотезы на основе скорректированного множества $\mathbf{T}_{train}$, а на *седьмом* вычисляются степени объясняющей способности новых гипотез.

Заключительный, *восьмой*, этап предназначен для объединения гипотез, порожденных всеми q потоками и упорядочивания их по степени объясняющей способности.

Таким образом, предлагаемый метод абдуктивного вывода, во-первых, позволяет ранжировать множество индуктивно сгенерированных гипотез по убыванию степени объясняющей способности, что значительно облегчает для исследователя задачу анализа порожденных гипотез; во-вторых, обнаруживает и исключает из обучающего множества шумовые тексты, не порождающие информативные гипотезы; в-третьих, снижает вероятность возникновения эффекта переобучения за счет использования процедуры перекрёстной проверки.

Результаты исследования и их обсуждение

Для подтверждения эффективности разработанного метода абдуктивного вывода были проведены эксперименты с коллекцией отзывов о фильмах семинара РОМИП-2011 [6]. При этом ставились две задачи:

1) определение подмножества гипотез с высокой степенью объясняющей способности;

2) установление зависимости качества анализа от степени коррекции обучающих данных.

В отзывах о фильмах были выделены предложения, составившие обучающее множество текстов. Анализ на уровне предложений позволяет добиться выражения единственной тональности в пределах одного текста, что является важным условием применения ДСМ-метода. Всего были получены 97850 предложений.

На первом этапе метода абдуктивного логического вывода исходное множество предложений было разбито на 5 блоков ($q = 5$), по 19570 предложений в каждом. При этом в обучающем множестве для каждого потока оказалось $4 \times 19570 = 78280$ предложений. В ходе выполнения второго этапа в среднем для каждого потока сгенерировано 57564 гипотезы. На третьем и четвертом этапах вычислялись соответственно степень объясняющей способности гипотез и степень значимости обучающих текстов. Для отбора предложений на пятом этапе использовалось следующее правило: оставлялись тексты с наибольшими значениями степеней значимости, сумма степеней значимости которых равна не менее 90 % от суммарной степени значимости всех текстов. В результате такого отбора в скорректированном обучающем множестве оказалось в среднем по всем потокам

23147 предложений (29,6% от 78280 исходных текстов). На шестом этапе на основе нового обучающего множества было порождено в среднем 18276 гипотез (31,8% от 57564 начальных гипотез). На седьмом этапе вычислялись степени объясняющей способности вновь порожденных гипотез.

На восьмом этапе гипотезы, сгенерированные во всех пяти потоках были объединены, в результате чего получился список, состоящий из 23569 гипотез. Данный список был упорядочен по степени объясняющей способности гипотез и проанализирован. Анализ списка позволяет сделать вывод о том, что на основе предложенного абдуктивного метода среди множества индуктивно порожденных гипотез выявляются гипотезы, адекватные предметной области. В таблице приведены примеры наборов слов (в нормальной форме), входящих в позитивные и негативные гипотезы.

Примеры позитивных и негативных гипотез

Позитивные гипотезы	Негативные гипотезы
1) хороший, добрый 2) пересматривать 3) добрый, милый 4) любить, настоящий 5) смотреть, удовольствие	1) фильм, претензия 2) зря 3) жалко, потратить 4) полный, бред 5) скучный, затянутый

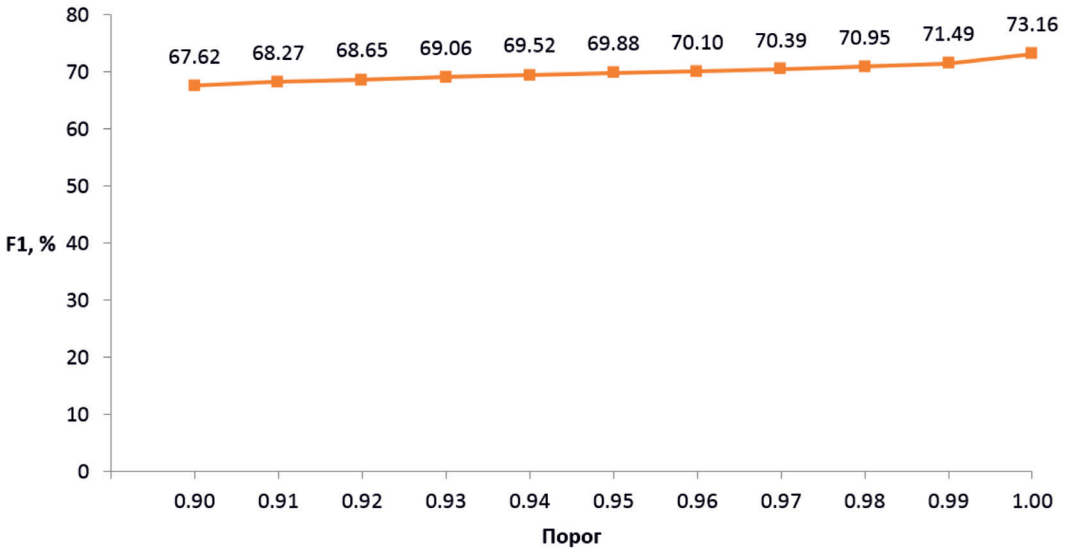

Рис. 2. Зависимость качества анализа тональности от степени коррекции обучающих данных

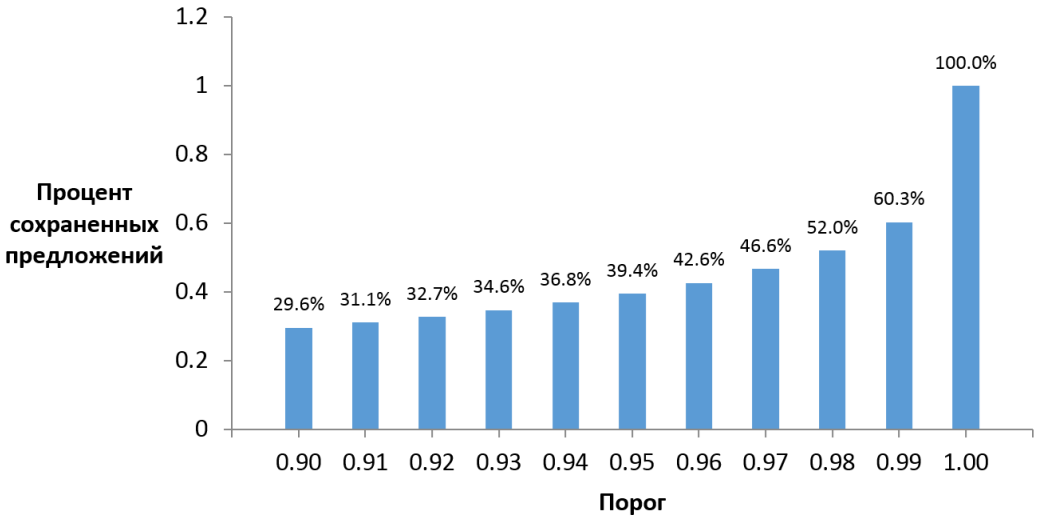

Рис. 3. Зависимость доли сохраняемых текстов от степени коррекции обучающих данных

Для решения второй задачи была построена зависимость качества анализа тональности от степени коррекции обучающих данных (рис. 2). На рис. 2 по оси абсцисс отложены пороговые значения, равные отношению суммы степеней значимости сохраняемых текстов к суммарной степени значимости всех текстов, а по оси ординат – значение F1-меры, которая представляет собой стандартную метрику качества анализа тональности [6].

На рис. 3 представлена зависимость процента сохраняемых текстов от степени коррекции обучающих данных.

Из рис. 2 и 3 можно сделать вывод, что всего 60,3 % предложений (порог 0,99) вносят наиболее существенный вклад в качество анализа (значение F1-меры отличается от максимального на 1,67 %). При этом остальные 39,7 % предложений можно исключить из обучающего множества, повысив за счет этого производительность обработки.

Заключение

В статье предложен метод абдуктивного логического вывода, позволяющий анализировать тональность коллекций текстов на основе ДСМ-метода за счет ранжирования множества порожденных гипотез по степени объясняющей способности, исключения из процесса анализа неинформативных текстов и снижения вероятности возникновения эффекта переобучения. Эксперименты, проведенные с применением коллекции отзывов о фильмах семинара РОМИП-2011, подтвердили эффективность разработанного метода.

В дальнейшем планируется провести детальное исследование характеристик множеств индуктивно порождаемых и абдуктивно принимаемых гипотез.

Работа выполнена в рамках государственного задания Минобрнауки РФ, проект № 586.

Список литературы

1. Котельников Е.В. Повышение быстродействия ДСМ-метода в задачах обработки текстовой информации // Труды Четырнадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2014 (24–27 октября 2014 года, г. Казань). – Казань: Изд-во РИЦ «Школа», 2014. – Т. 2.– С. 274–282.

2. Котельников Е.В. Функция оценки информативности гипотез для анализа тональности текстов на основе ДСМ-метода // Фундаментальные исследования. – 2014. – № 11(10). – С. 2150–2154.

3. Котельников Е.В. Классификация отзывов о фильмах с использованием ДСМ-метода // В мире научных открытий. – 2013. – № 6.1 (42). – С. 225–242.

4. Финн В.К. Эпистемологические основания ДСМ-метода. Ч. I // НТИ. Сер. 2. Информационные процессы и системы. – 2013. – № 9. – С. 1–29.

5. Abductive inference: Computation, philosophy, technology / Josephson J., Josephson S. (Eds.). New York: Cambridge University Press, 1996.

6. Chetviorkin I., Braslavskiy P., Loukachevitch N. Sentiment Analysis Track at ROMIP 2011 // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue». – 2012. – № 11(18). – Vol. 2. – P. 1–14.

7. Kakas A.C. Abduction // In Encyclopedia of Machine Learning / C. Sammut and G.I. Webb (eds.), Springer, 2012. – P. 3–9.

8. Kakas A.C., Kowalski R., Toni F. Abductive logic programming // Journal of Logic and Computation. – 1992. – Vol. 2(6). – P. 719–770.

9. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection // Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. – 1995. – № 2(12). – P. 1137–1143.

10. Kraus M., Charles S. Peirce's Theory of Abduction and the Aristotelian Enthymeme From Signs // Anyone Who Has a View, Argumentation Library. – 2003. – Vol. 8. – P. 237–254.

11. Peirce C.S. Philosophical writings / Ed. J. Buchler. N. Y: Dover Publ. Co., 1955.

12. Strabykin D.A. Logical Method for Predicting Situation Development Based on Abductive Inference // Journal of Computer and Systems Sciences International. – 2013. – Vol. 52(5). – P. 759–763.

References

1. Kotelnikov E.V. *Trudy Chetyrnadcatoj nacional'noj konferencii po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2014* [Proceedings of the 14th national conference on artificial intelligence with international participation]. Kazan, School, 2014, T. 2, pp. 274–282.

2. Kotelnikov E.V. *Fundamental'nye issledovanija*, 2014, no. 11(10), pp. 2150–2154.

3. Kotelnikov E.V. *V mire nauchnyh otkrytij*, 2013, no. 6.1(42), pp. 225–242.

4. Finn V.K. *NTI. Ser. 2. Informacionnye processy i sistemy*, 2013, no. 9, pp. 1–29.

5. Abductive inference: Computation, philosophy, technology / Josephson J., Josephson S. (Eds.). New York: Cambridge University Press, 1996.

6. Chetviorkin I., Braslavskiy P., Loukachevitch N. *Annual International Conference «Dialogue»*, 2012, no. 11(18), Vol. 2, pp. 1–14.

7. Kakas A.C. *Encyclopedia of Machine Learning* / C. Sammut and G.I. Webb (eds.), Springer, 2012, pp. 3–9.

8. Kakas A.C., Kowalski R., Toni F. *Journal of Logic and Computation*, 1992, Vol. 2(6), pp. 719–770.

9. Kohavi R. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1995, no. 2(12), pp. 1137–1143.

10. Kraus M., Charles S. *Anyone Who Has a View*, Argumentation Library, 2003, Vol. 8, pp. 237–254.

11. Peirce C.S. *Philosophical writings* / Ed. J. Buchler. N.Y: Dover Publ. Co., 1955.

12. Strabykin D.A. *Journal of Computer and Systems Sciences International*, 2013, Vol. 52(5), pp. 759–763.

Рецензенты:

Страбыкин Д.А., д.т.н., профессор, заведующий кафедрой электронных вычислительных машин, Вятский государственный университет, г. Киров;

Прозоров Д.Е., д.т.н., профессор кафедры радиоэлектронных средств, Вятский государственный университет, г. Киров.

Работа поступила в редакцию 01.04.2015.