Automatic Data Quality Evaluation for Text Classification

Jiazheng Li
Key Lab of Intelligent Information Processing, Institute of Computing Technology
University of Chinese Academy of Sciences, Beijing, China
lee_jiazh@163.com

Abstract

Data quality is critical for machine learning, but its evaluation usually relies on the performance of used models. A model-independent data quality evaluation metric is needed. This paper proposes a convenient metric called DQTC to quantify the data quality for text classification based on information theory. And an experiment is conducted to verify the relevance between DQTC and model performance. Finally, we describe the linguistic improvement that should be considered. The code is available online [1].

1 Introduction

Data quality has been studied for decades, and previous researches mainly focus on the improvement of data, the management of data [1], etc. The data quality is usually evaluated by the feedback from models, users and data annotators, which relies the external resources and defers the insight of data. Existing data-centric benchmarks such as DataCLUE [2] still adopt a fixed model to train with the modified data and use the performance as the data quality [6] . However, selecting different models may involve bias to the data quality evaluation. Similar to the metrics that evaluate the performance of models for different learning tasks (e.g., ROUGE for summarization), a standard and generic data quality evaluation metric without introducing models is necessary for data-centric research.

In this paper, we design a simple data quality evaluation metric DQTC to evaluate the data provided for text classification based on information theory. And an experiment is conducted on IMDB movie reviews (a data set for sentimental analysis), where the data is processed by different text preprocessing methods to verify the relevance between DQTC and model performance.

2 Data Quality for Text Classification

For text classification, a better data set should have more balanced samples and more important words, where the important words are the words that have relatively different occurrences in the samples from different categories. For example, if a word w appears in category c_1 much more than category c_2, then w is a strong feature to distinguish the category of a given sample. A word's importance (or significance) can be calculated by the statistical term weighting methods, such as term frequency(tf), inverse document frequency (idf), mutual information, chi square test, etc.

The computation of the evaluation metric **Data Quality for Text Classification** (DQTC) is shown in formula 1, where W is the vocabulary of given corpus, $S(w)$ is a function that returns the weight of word w, C is the category set, $|c|$ is the sample count and $|\bar{C}|$ is the average sample count for all categories.

[1] https://gist.github.com/gajanlee/7faac80c2ea9cd3032c53d3079059c6e
[2] https://github.com/CLUEbenchmark/DataCLUE

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Table 1: DQTC and the accuracy by different models

		Model				
Operation	DQTC	NB	SVM	LightGBM	FastText	KNN
Original	3.45	0.787	0.521	0.874	0.779	0.606
wo Stop	3.50	0.783	0.524	0.863	0.827	0.570
LowerCase	3.92	0.786	0.519	0.875	0.783	0.607
Lemma	4.00	0.788	0.520	0.874	0.780	0.615
f20	4.04	0.787	0.521	0.873	0.816	0.603
chi20	4.05	0.787	0.525	0.873	0.809	0.601
Stem	4.39	0.786	0.520	0.873	0.813	0.618
chi40	4.84	0.789	0.525	0.874	0.815	0.601

$$DQTC = \frac{\sum_{w \in W} S(w)}{|W| \times (\sum_{c \in C} |c| - |\bar{C}| + 1)} \tag{1}$$

DQTC mainly evaluates the data set from the two aspects: 1) **Balance.** If a data set is entirely balanced, then DQTC is the sum of all words' weights; or the sum of the sample count differences between each category and average count is as a punished factor to reduce the DQTC. 2) **Conciseness.** If the size of vocabulary W is fixed, and more important words means the sum score is higher, then DQTC is higher, namely, less the words with lower significance.

3 Experiment

Preprocessing is usually the first step in the pipelines of machine learning, which can make the corpus clean to improve the performance of models. The common methods mask and remove the irrelevant content. We extend the methods in [3] to produce different data sets, include **Original**, apply no operations; **wo Stop**, remove all stop words in the texts; **Lower Case**, the texts are all in lower case; **Stem** and **Lemma**, use the NLTK package to stem and lemmatize the words respectively; **f20**, use F-test as feature selection and remove the last 20 percents words; **chi20** and **chi40** use chi square as feature selection and remove the last 20 and 40 percent words respectively. The used machine learning models include **NB**, Naive Bayes classifier; **SVM**; **LightGBM** [4]; **FastText** [2] and **KNN**. The models cover the random forest, clustered and neural classifiers to help verify the data quality.

We use chi square as S function to evaluate the DQTC of the data sets. The used data set is collected from IMDB movie reviews for sentimental analysis [5], where the train and test data both have 25000 samples. We use the two categories *pos* and *neg* for text classification, and each category contains 12500 samples, namely, the data set is balanced. The experimental result is shown in Table 1.

Generally, we can observe that the DQTC grows in direct proportion with the accuracy, and the preprocessing methods affect weakly about the performance. DQTC orients the machine-readable data and overlooks the readability. Consider the following three processed sentences, the *wo Stop* and *Stem* adopts incomplete sentence as train data, which trades off higher DQTC and better performance against text coherence. The DQTC should consider more linguistic requirements.

- **Original:** Starts out with a opening scene that is a terrific example of absurd comedy
- **wo Stop:** Starts scene terrific absurd comedy
- **Stem:** Start out with a open scene that is a terrif exampl of absurd comedi

4 Conclusion

The experimental result shows that DQTC can provide an insight of data quality generally, but more linguistic features should be considered. And we expect the investigation about the data quality metrics can accelerate the development of data-centric benchmark.

References

[1] Venkat Gudivada, Amy Apon, and Junhua Ding. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, 10(1):1–20, 2017.

[2] Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, 2017.

[3] Subbu Kannan, Vairaprakash Gurusamy, S Vijayarani, J Ilamathi, M Nithya, S Kannan, and V Gurusamy. Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, 5(1):7–16, 2014.

[4] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

[5] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1015`.

[6] Liang Xu, Jiacheng Liu, Xiang Pan, Xiaojing Lu, and Xiaofeng Hou. Dataclue: A benchmark suite for data-centric nlp, 2021.