ИЗВЛЕЧЕНИЕ ЗНАНИЙ ИЗ ЕСТЕСТВЕННЫХ ТЕКСТОВЫХ ДАННЫХ

П. А. Филиппович

Работа с естественными текстовыми данными – нетривиальная задача по многим причинам. Необходимо учитывать синтаксические и семантические особенности как языка, так и автора, а также смысл отдельных слов и текста в общем. При создании моделей для работы с текстом добавляется еще одна задача: необходимо представить текст в виде, который позволит работать с ним инструментам анализа. [1, 2].

Рассмотрим задачу предсказания величины оплаты труда по описанию вакансии (задача регрессии).

Есть набор данных с вакансиями, в котором для каждой записи присутствует следующая информация: табл.1.

Поля SalaryNormalized и SalaryRaw отсутствуют в тестовой выборке. Значения из SalaryNormalized являются искомыми скрытыми знаниями, которые необходимо извлечь из тестовой выборки.

В обучающей и тестовой выборках по 10000 записей. Модель оценивается метрикой средней абсолютной ошибки.

Таблица 1

Доступная информация о вакансиях

Название поля	Тип данных	Описание
Title	текст	Название вакансии
FullDescription	**текст**	Полное описание вакансии с пропусками на месте зарплаты
LocationRaw	текст	Полное географическое местоположение
LocationNormalized	текст	Город
ContractType	текст	Тип рабочего дня
ContractTime	текст	Срок трудоустройства
Company	текст	Название компании
Category	текст	Сфера деятельности
SourceName	текст	Сайт, на котором была найдена вакансия
SalaryRaw	текст	Зарплата в виде, указанном в описании вакансии
SalaryNormalized	целое число	Зарплата, приведенная к целому числу единого формата

В рамках статистического анализа проверялись соотношения длин слов, длин описания вакансий, длин названий вакансий и их различные агрегации между обучающей и тестовой выборками.

В тестовой выборке присутствует из обучающей выборки:

- 48 % компаний

- 67 % городов
- 100 % категорий вакансий

Распределение зарплат по сферам деятельности, городу расположения и типу рабочего дня позволило выявить определённые тенденции по каждому критерию в отдельности, но с некоторым количеством статистических выбросов.

Проведённый статистический анализ позволил предположить, что модель, построенная на обучающей выборке с использованием всех доступных полей, может достоверно предсказывать целевую переменную для тестовой выборки.

Было добавлено поле Gathered, которое представляет собой объединение всех полей, доступных в тестовой выборке, через пробел. К данным применялись следующие комбинации подходов к предобработке текстовых данных:

- Bag of Words.
- Взвешивание tf-idf [3].
- Стемминг + Bag of Words.
- Стемминг + взвешивание tf-idf.
- Word2vec для каждого термина в отдельности [4].
- Word2vec с усреднением всех векторов терминов (для нейросетевых моделей) [4].

Используемые модели:

- Linear: линейная регрессия, самая простая линейная модель. Использовалась в качестве базового решения для сравнения результатов остальных построенных моделей.
- Huber: устойчивая регрессия с функцией потерь Хьюбера, линейная модель, которая может оптимизировать либо функцию квадратичной ошибки, либо функцию абсолютной ошибки. Была выбрана из-за метрики оценивания данного задания [5].
- RFR: ансамблевая модель случайных решающих деревьев. Отличается повышенной интерпретируемостью [6].
- GBR: градиентный бустинг решающих деревьев. Отличается последовательным построением слабых решающих деревьев [7].
- Нейросетевые модели: свёрточная сеть [8] (№1), свёрточная сеть [8] с плотным слоем (№2), сети из 2 (№1) и 3 (№2) плотных слоёв.

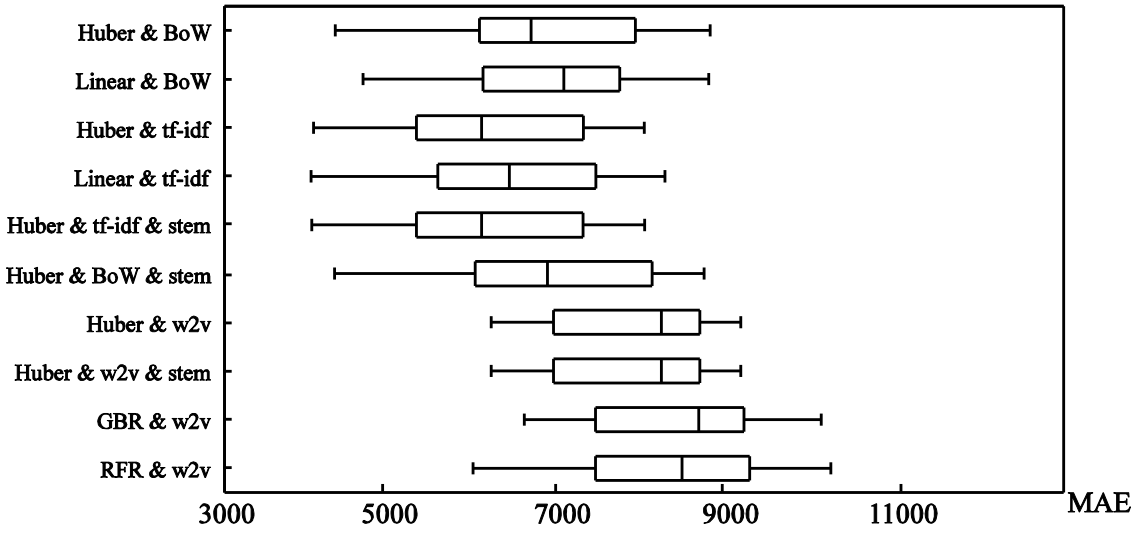

Рис. 1. Результаты линейных и ансамблевых моделей

РЕЗУЛЬТАТЫ

Сперва проверили и подтвердили предположение, что использование в моделях комбинации всех доступных полей (поле Gathered) даст лучший результат. Наилучший результат (рис.1) показала модель с HuberRegressor, стеммингом и tf-idf взвешиванием терминов. Подобный результат можно обусловить тем, что HuberRegressor с выбранными параметрами оптимизирует функцию абсолютной ошибки, при этом метрика для оценивания модели – средняя абсолютная ошибка. Tf-idf взвешивание позволяет учитывать частоту термина не только в рамках одной записи, но и во всём множестве записей. Благодаря стеммингу улучшается обобщающая способность модели для tf-idf взвешивания.

Из табл. 2 видна повышенная способность нейросетевых моделей переобучаться на обучающем множестве. При этом наилучший результат обучения данных моделей не превосходит результата, достигнутого Huber со стеммингом и tf-idf взвешиванием в рамках данной задачи.

Таблица 2

Результаты обучения нейросетевых моделей

Архитектура сети	Валидация на	
	обучающей выборке	тестовой выборке
Свёрточная сеть №1	8863	10180
Свёрточная сеть №2	6172	10360
Полносвязная сеть №1	8237	9120
Полносвязная сеть №2	7364	8470

ЗАКЛЮЧЕНИЕ

Заметим, что построенная модель обладает средней абсолютной ошибкой ~20% от среднего значения целевой переменной при кросс-валидации. Учитывая сильный разброс целевой переменной (от 5 т до 175 т), данная модель может использоваться на практике для оценки соответствия оплаты труда для вакансий, а также для грубой оценки резюме.

В дальнейшем, развивая полученную модель, можно будет генерировать описания вакансий по заданным критериям и величине оплаты либо использовать в рекомендательных системах на бирже труда.

В рамках поставленной задачи также можно использовать: иные комбинации полей записей; специальным образом обрабатывать записи, значения целевой переменной которых являются статистическими выбросами; предсказывать измененную целевую переменную (например, прологарифмировав) либо же разбить целевую переменную по классам и перейти от задачи регрессии к задаче классификации. При внесении изменений в выборку данных нужно вносить изменения в модель, зачастую переделывать модель, используя новые подходы.

Литература

1. *Kurbatow, A.* The research of text preprocessing effect on text documents classification efficiency / A. Kurbatow // "Stability and Control Processes" in Memory of V.I. Zubov (SCP), 2015 International Conference / IEEE – 2015.
2. *Madsen, R.E.* Pruning the vocabulary for better context recognition / R.E. Madsen, S. Sigurdsson, L.K. Hansen, J. Larsen // Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on / IEEE – 2004.
3. *Nakashima, T.* Daily clustering for the electronic newspaper based on the analysis of trends / T. Nakashima, R. Nakamura // Pacific Rim Conference on Communications, Computers and Signal Processing, 1999. Conference Proceedings, Victoria, BC / IEEE - 1999, pp. 51–54.
4. *Ma, L.* Using Word2Vec to process big text data / L. Ma, Y. Zhang // Big Data (Big Data), 2015 IEEE International Conference on / IEEE – 2015.
5. *Mangasarian, O.L.* Robust linear and support vector regression / O.L. Mangasarian, D.R. Musicant // in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 9 pp. 950–955 / IEEE - Sep 2000.
6. *Tin Kam Ho* The random subspace method for constructing decision forests / Tin Kam Ho // in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832–844 / IEEE - Aug 1998.
7. *Polikar, R.* Ensemble based systems in decision making / R. Polikar // in IEEE Circuits and Systems Magazine, vol. 6, no. 3, pp. 21-45 / IEEE - Third Quarter 2006.
8. *Kim, Y.* Convolutional neural networks for sentence classification / Yoon Kim // EMNLP 2014.