达观数据
DATA GRAND

“达观杯”文本智能处理挑战赛

TNT_000_

DC竞赛
www.dcjingsai.com

目录

CONTENT

团队介绍

成员介绍

杨亚涛-中山大学
InplusLab

罗志鹏-北京大学
Microsoft

肖小粤-中山大学
InplusLab

何嘉伟
湖南大学

丁晓菲
湖南大学

团队成绩

天池大数据	CIKM CUP 2018	1st/1027
KDD CUP 2018	Last 10-Day Prediction	1st /4000+
KDD CUP 2018	Second 24-Hour Prediction	1st /4000+
DataCastle	微博传播热度预测大赛	1st/668
Kesci	BOT2016人工智能聊天机器人大赛	1st /78
KDD CUP 2018	Main Track	2nd/4000+
DataFountain	2017CCF 让AI当法官	2nd /415
DataFountain	唯品会用户购买行为预测	2nd/542
Kesci	拍拍贷魔镜算法大赛	2nd/485
DataCastle	微额贷款人品预测大赛	3rd/1666
DataFountain	2016CCF 搜狗用户画像	3rd/894
DataFountain	2016CCF 客户用电异常行为分析	4th/888
DataFountain	2017CCF 360人机大战	5th/888
KDD CUP 2017	Tasks1	6th/3582
滴滴研究院	第一届滴滴算法大赛	8th/7000+
天池大数据	阿里聚安全算法挑战赛	9th/1124

赛题分析

赛题任务

任务：通过长文本的字和词的序列数据，预测文本类别

词(word_seg)	字(article)
816903 597526 520477	7368 1252069 365865
1179558 1033823 758724	755561 1044285 129532
632718 422098 520477	105340 856005 1010033
414956 625597 1203094	679500 426236 755561
441513 596474 660569	426236 697742 1044285
995362 924085 1278762	678940 1007781 1205281
1155376 1109510 …	263861 195449 …

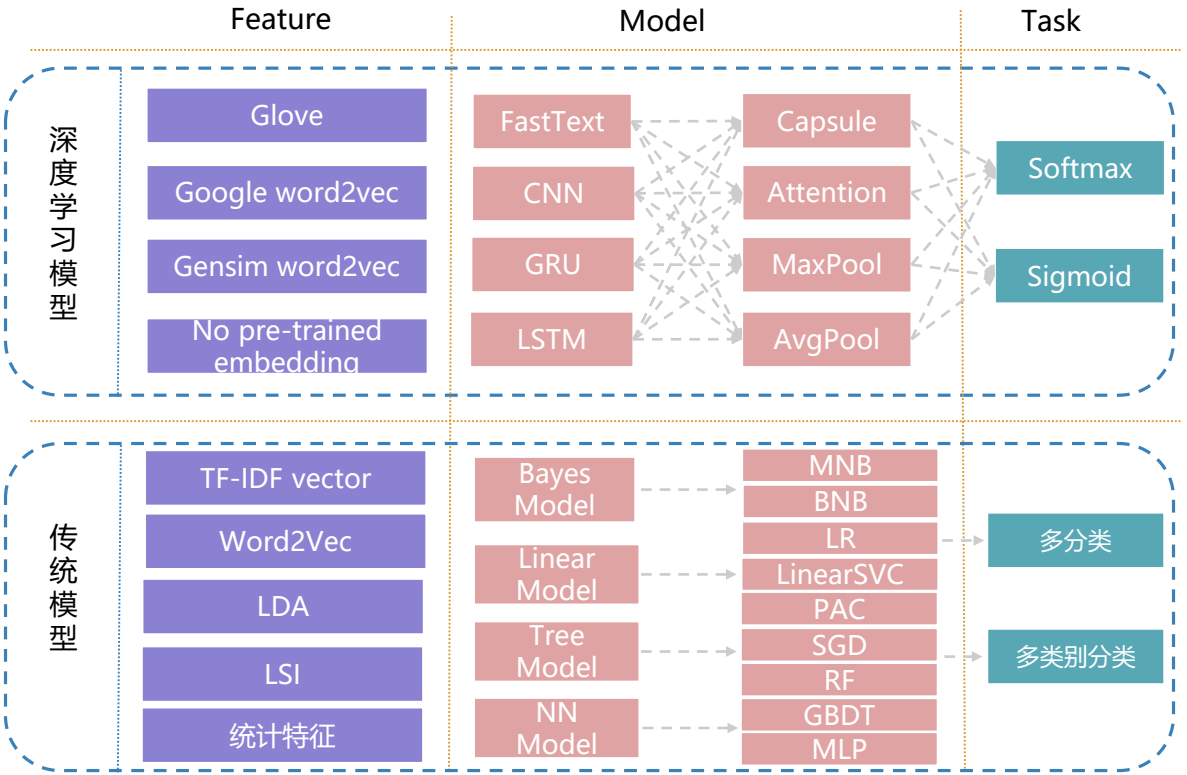
预测 → **类别(class)**
1~19

评价指标

Macro-F1： $\langle F1 \rangle = \frac{1}{n}\sum_{i}^{n} F1_i = \frac{1}{n}\sum_{i}^{n} \frac{2 \cdot P_i \cdot R_i}{P_i + R_i}$

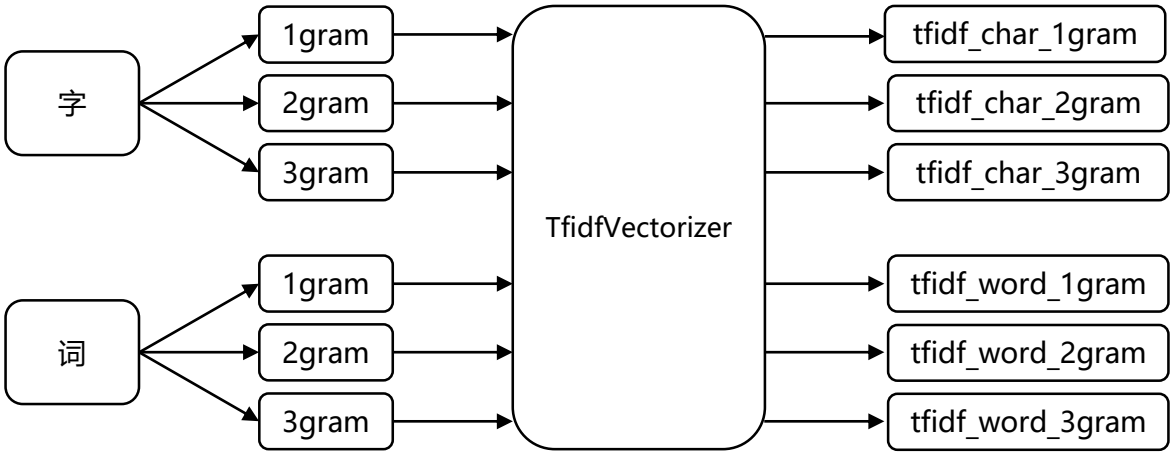
总体框架

模型架构

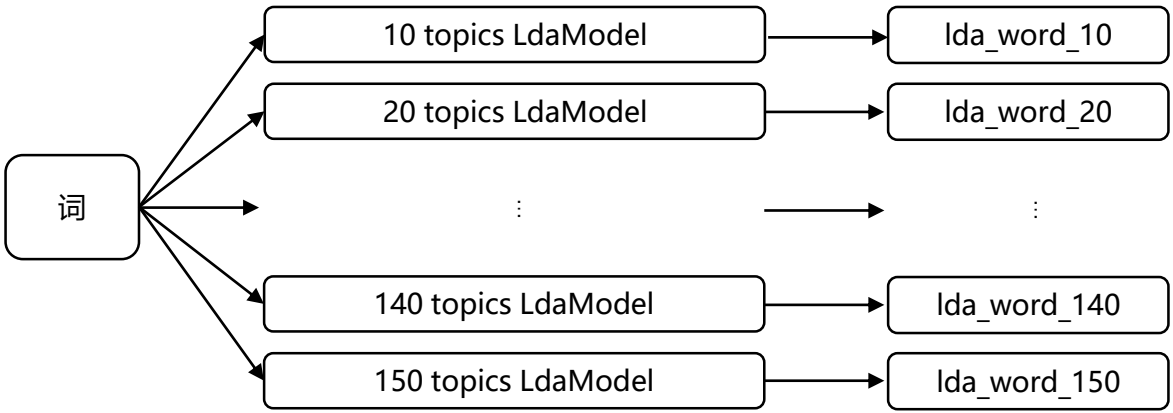

Feature
Model
Task
深度学习模型
Glove
Google word2vec
Gensim word2vec
No pre-trained embedding
FastText
CNN
GRU
LSTM
Capsule
Attention
MaxPool
AvgPool
Softmax
Sigmoid
传统模型
TF-IDF vector
Word2Vec
LDA
LSI
统计特征
Bayes Model
Linear Model
Tree Model
NN Model
MNB
BNB
LR
LinearSVC
PAC
SGD
RF
GBDT
MLP
多分类
多类别分类

特征构建

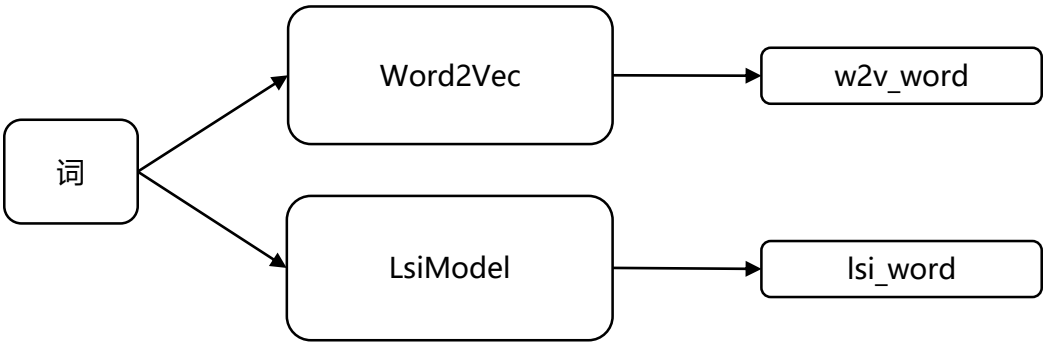
TF-IDF特征


字
词
1gram
2gram
3gram
1gram
2gram
3gram
TfidfVectorizer
tfidf_char_1gram
tfidf_char_2gram
tfidf_char_3gram
tfidf_word_1gram
tfidf_word_2gram
tfidf_word_3gram

LDA特征


词
10 topics LdaModel
20 topics LdaModel
140 topics LdaModel
150 topics LdaModel
lda_word_10
lda_word_20
lda_word_140
lda_word_150

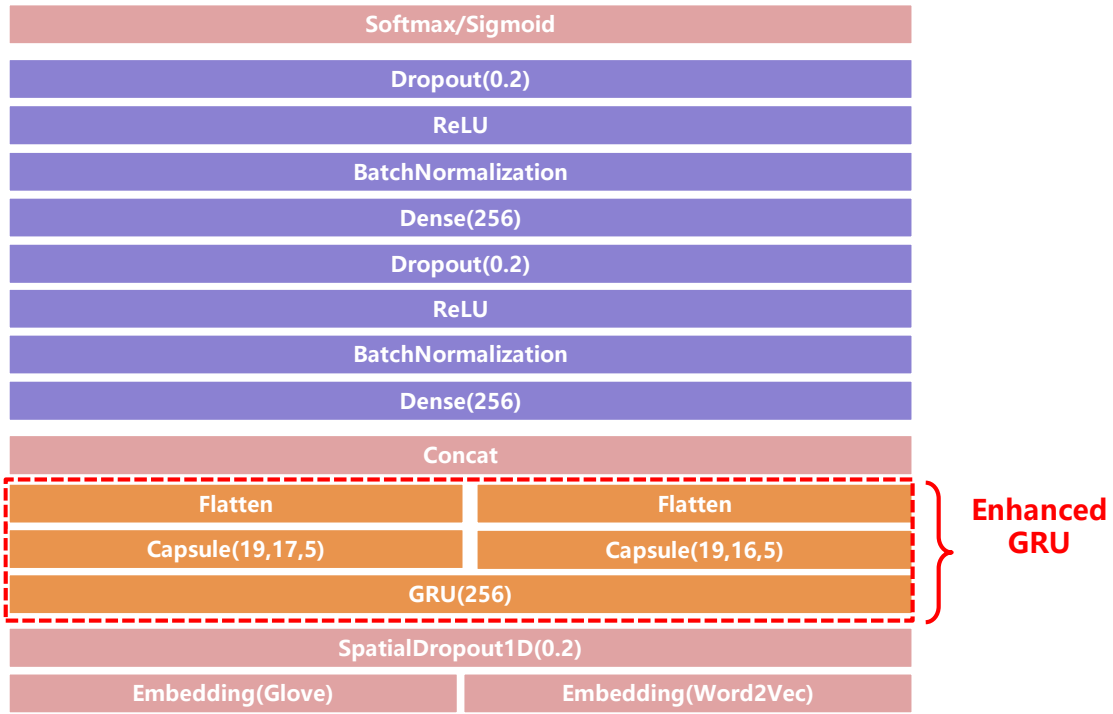
LSI特征+Word2Vec特征



模型构建

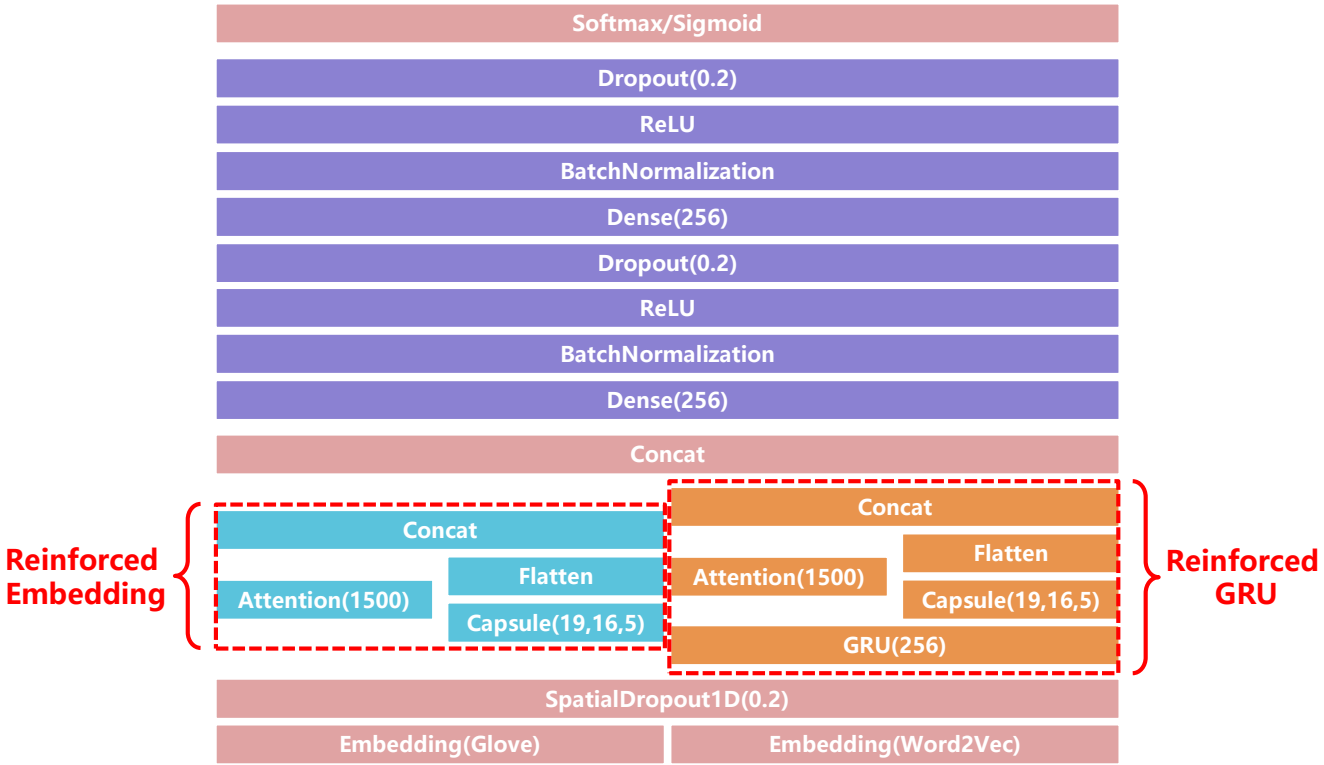
Hybrid NN-1

0.8013



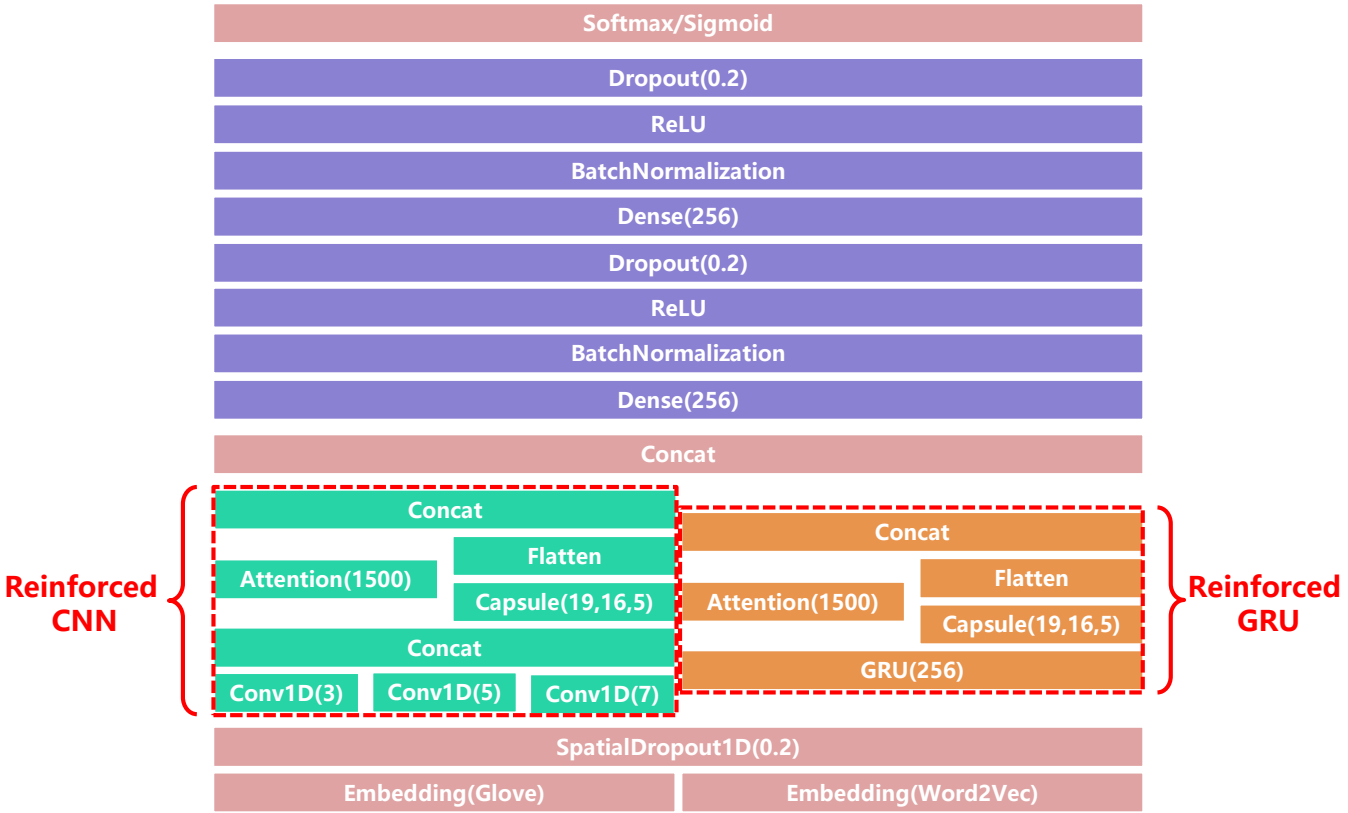
Hybrid NN-2

0.8017



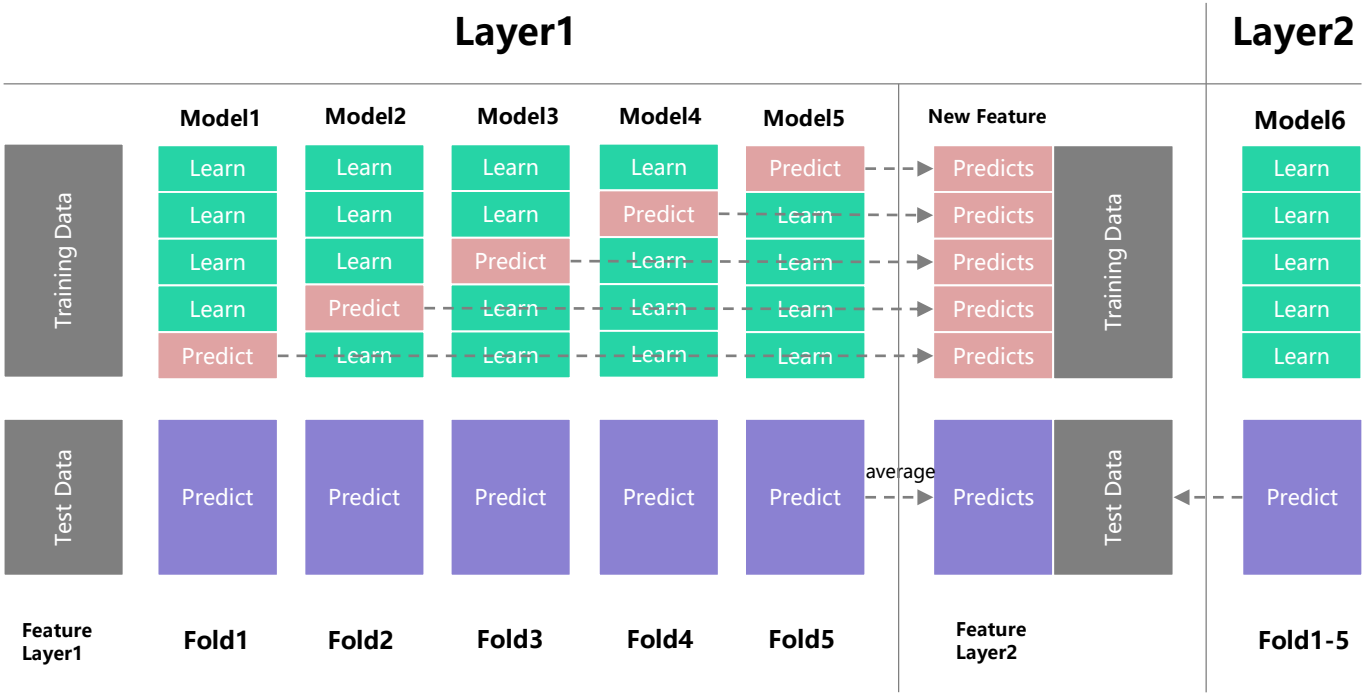
Hybrid NN-3

0.8019

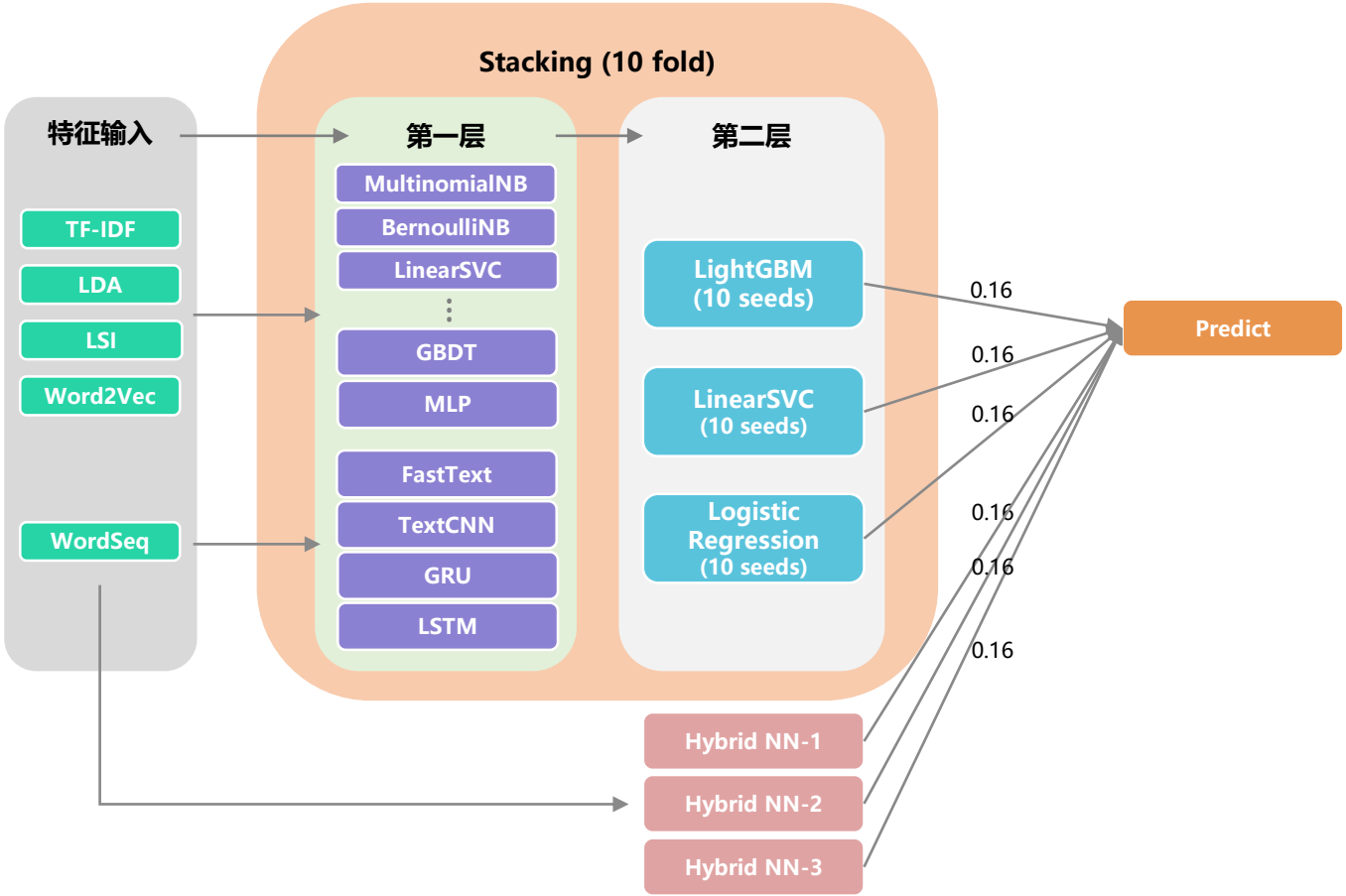


模型融合

Stacking框架


Layer1
Layer2
Model1
Model2
Model3
Model4
Model5
New Feature
Model6
Training Data
Learn
Predict
Predicts
Test Data
average
Feature Layer1
Fold1
Fold2
Fold3
Fold4
Fold5
Feature Layer2
Fold1-5
10 folds
10 seeds avg

融合策略


Stacking (10 fold)
特征输入
TF-IDF
LDA
LSI
Word2Vec
WordSeq
第一层
MultinomialNB
BernoulliNB
LinearSVC
GBDT
MLP
FastText
TextCNN
GRU
LSTM
第二层
LightGBM (10 seeds)
LinearSVC (10 seeds)
Logistic Regression (10 seeds)
Hybrid NN-1
Hybrid NN-2
Hybrid NN-3
0.16
0.16
0.16
0.16
0.16
0.16
Predict

总结展望

总结

- 网络结构的创新和改进对本赛题的效果是明显的
- 预训练的Embedding能加快网络的训练，并且效果俱佳
- 传统模型对于融合的提升是巨大的

展望

- 尝试其他的网络结构，例如：DPCNN，HAN等
- 混合网络的参数调优
- 融合系数的优化

Thanks!

TNT_000_