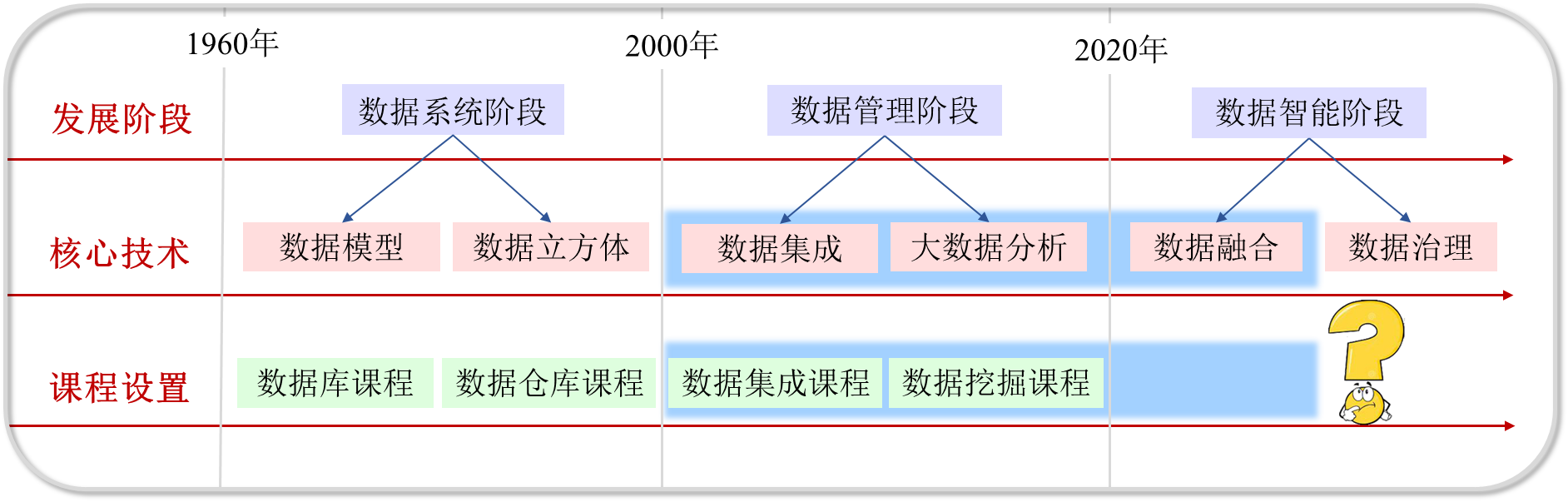
数据管理与数据智能导论

前言**（郝新丽）**

当前全球经济数字化转型不断加快，提升我国公民的数据素养具有重要意义。对于高校大学生乃至科研人员而言，数据素养或将成为重要的科学素养之一。本团队依据多年的研究成果，将数据融合和数据智能作为建立高校大学生数据素养的有效途径，培养学生自主地从数据中发现价值的能力。另外，本课程探索一种新型的“沉浸式代入教学模式”，通过制定自组织的课程内容、自主性的教学模式、自适应的评价体系，尝试解决学生日益增长的对社会技术加速变革中新生长知识的需求与传统课程体系所固定的知识内容之间的矛盾，以及学生日益丰富的获取信息的途径与传统相对单一的“灌输式”教学方式之间的矛盾。

**一、课程介绍（郝新丽）**

在数字化转型时代，提高公民的数据素养（digital literacy）具有重要意义。2021年11月，中央网络安全和信息化委员会印发了《提升全民数字素养与技能行动纲要》，明确了提升全民数字素养与技能的指导思想、基本原则、发展目标等。数据素养强调处理、分析和解读数据的能力，对于高校大学生乃至科研人员而言，数据素养将成为与数学素养同等重要的科学素养之一。



数据素养包括对数据进行收集、组织、分析、挖掘价值、辅助决策等基本内涵，与数据管理的发展历程相呼应。如上图所示，数据管理的发展历程包括数据系统、数据管理、数据智能3个阶段：数据系统最初应企业经营组织管理的需求而生，其重点在于数据的组织和效率，核心技术主要包括数据模型和数据立方体，数据库和数据仓库两门课程涵盖了此阶段的核心知识；互联网大数据应用兴起改变了数据产生方式，因而进入了数据管理阶段，其重点在于数据集成和数据分析，主要由数据集成与数据挖掘两门课程进行教学；当下面临数字化转型的巨大应用，数据管理将让位于数据智能，其重点在于数据融合和数据治理，尚未产生新的课程与之呼应。

本课程依据本团队多年的研究成果，跨越数据管理与数据智能2个阶段，围绕5个核心知识板块：数据对齐、数据学习、数据增强、实时分析和知识图谱，弥补现有课程体系的缺失，培养学生从数据中发现价值的能力，从而建立数据素养。

二、教学方法（郝新丽）

在数字化转型时代，知识产生方式与获取方式都发生了改变，我们传承已久的人才培养模式面临着如下两点挑战：

首先，学生日益增长的对社会技术加速变革中新生长知识的需求与传统课程体系所固定的知识内容之间产生矛盾。传统的课程体系将发展成熟的知识组织为相对固定的知识内容，转化的过程虽然严谨，但相对缓慢。而在数字化转型的时代，知识更迭的速度加快，传统的知识组织方式已难以满足当代学生对社会技术加速变革中新生长知识的需求。因此需要一种能够将知识迅速转化为课程内容的动态教学方式。

其次，学生日益丰富的获取信息的途径与传统相对单一的“灌输式”教学方式之间产生矛盾。传统的教学方法为“灌输式”教学，极大地依赖于课堂讲授，由教师单方面输出，学生被动地接受。而在数字化转型的时代，学生获取知识与信息的途径随之多元化。互联网上充斥着众多的专业博客、学术讲座、开源代码等等，极大地扩展了学生自主学习的渠道，超越了传统课堂的时间与空间限制。

面对这样的发展趋势，为了尽快提升学生的数据素养，教师的教学的方法和形式也应与时俱进，探讨新的课程组织模式和教学模式，避免故步自封、照本宣科。据此，我们做出初步尝试，提出一种沉浸式代入教学模式，培养学生从数据中自主发现问题、解决问题、挖掘价值的能力。该方式具体包括以下三个特征：

1）自组织的课程内容。在本课程中，教师提供相对固定的主题内容，始终围绕数据融合与数据智能所涉及的5大知识板块，学习者可以依据主题自主选择阅读文献。这在一定程度上解决了学生对于前沿知识的需求。

2）自主性的教学模式。本课程通过具体的计算任务将学生代入到数据融合与数据智能的知识体系，促使学生自主地寻找资源，沉浸式地感受困难与挑战所在，激励学生在实践过程中形成独特的理解，从而培养学生自主解决问题的能力。

3）自适应的评价体系。考试为一种相对客观的考核方式，具有明确的标准，但学生自由发挥的空间小；实验、论文与报告等为一种相对主观的考核方式，学生可充分发挥主观能动性，但缺乏统一的量化标准，仁者见仁智者见智。本课程采用二者相结合的方式：每一个计算任务都具有定量的评价指标，可以量化解决方案的有效性；同时通过报告等方式为学生提供自由发挥、交流讨论的空间，另外，通过学生提交实验结果的次数来量化学生的努力程度。自适应的评价体系可以根据更为真实全面地评判学生的学习效果。

具体地，针对每个任务，授课教师提供初步的解决方案（baseline）、参考书籍和相关文献；同学们通过阅读书籍和文献，思考如何将其应用于解决当前任务；助教组织学生们课堂汇报交流自己的经验与收获，从而达到互相分享经验的目标。另外，要求同学们动手编程实践，切实解决研究案例中的问题，提交实验结果，并实时展示实验效果排行榜，激励学生保质保量地完成案例任务。

本课程所提出的基于计算任务的沉浸式代入教学与经典的案例教学有相似之处，同时也有显著区别，具体表现在如下三点：

1）本课程使用计算任务引导学生进行实践与思考，计算任务与法律或商务案例具有本质区别；

2）本课程在开放式讨论的基础上，要求学生用代码切实完成任务要求；

3）本课程的代入式教学具有定量的评判标准：算法效果排行榜，可以有效地激励学生不断探索更为有效的解决方案。

值得注意的是，本课程具有一定的局限性：“沉浸式代入教学模式”并不通用，只适用于侧重实践、同时计算任务丰富的课程。其无法替代传统的经典课程，仅作为对于现有课程体系的有效补充与完善。

**三、教学内容与要点**

1. 数据对齐（但唐朋）数据管理中数据集成下所面临的最根本的问题（与上文相呼应）

数据对齐旨在从不同结构，不同主题的数据库中识别与挖掘其中包含的“相同”实体对象，从而以此为依据进行更深入的挖掘研究。数据对齐广泛存在于数据集成场景中：（1）剔除合并数据库的重复元组；(2)不同语料集中的实体匹配；(3)相似度挖掘。本课程中我们以中英文学者对齐这一计算任务为导向，引导学生进入课程，启发同学通过特征工程、机器学习等基本方法、来实现目标，提升对数据的分析理解能力，并最终能够探索利用数据对齐来解决数据融合中所面临的相关问题。

1.**1 计算任务:中**英文学者对齐

许多中国学者既发表中文论文，也发表英文论文。为了调研国内学者在中外期刊上的发文数量，我们需要将以两种语言写成的论文的作者对应起来。然而，学者在用不同语言发表论文时署名的形式不同，比如，“刘伟”和“WeiLiu”，这导致不同学者的英文署名可能是相同的，“WeiLiu”可能是学者“刘伟”，也可能是学者“柳维”。本案例的任务是将多个看上去相互对应的中文名和英文名进行匹配和对齐，准确地寻找到中英文名字中真正对应的同一位学者。

（1）任务目标：根据中英文论文发表情况表，建立特征模型，挖掘数据集中的潜在信息，从中、英文确定并对齐相同的作者。

（2）基础模型：本次任务提供了特征工程中的常用模型词频-逆文档频率模型（TF-IDF），帮助同学对任务认识和理解。

（3）结果测评：本次任务提供了测试平台，并提供了基准代码（baseline），评价结果为：0.41。在此基础上进行改进并提交结果，系统将所有学生提交结果给出排名（展示基准代码的代码片段、排名截图）。目前课程中已经获得的最高分为：0.65。

|  |
| --- |
| Baseline |
| def getDpaper(mappingtab,papers,authorId):      if '-' not in authorId:          authorId=eval(authorId)      if authorId not in mappingtab:          return []      engPmap=mappingtab[authorId]      tpp=[]      for enp in engPmap:          pid=enp['paperId']          if pid in papers:              tpp.append(papers[pid])      return tpp  def get\_ch\_feature(cngPaper,keys):      AuthorsPname=[]      for ppp in cngPaper:          cons=[]          for pp in ppp:              valu=pp[keys]              if keys=='authors':                  valu=lazy\_pinyin(valu)              cons.append(str(valu))          AuthorsPname.append(' '.join(cons))      return AuthorsPname  def get\_en\_feature(engPaper,keys):      AuthorsPname=[]      ss=''      for ppp in engPaper:          valu=ppp[keys]          ss+=' '+(str(valu)).lower()      AuthorsPname.append(ss)      return AuthorsPname |



（4）友情提示：（有哪些建议和方法可以提升分值，3-5个，每个建议大概可以提升多少，列出点，最好给出参考分值。也不能太详细）

1.选用新的词向量模型Word2vec来完成特征工程,可以提升0.1左右的效果；2.从数据特征角度出发，挑选并组合不同特征计算相似度，可以提升0.1左右的效果；3.计算中英文论文发文时间的相似度，可以提升0.3左右的效果。

为进一步提升学生的数据分析处理能力，本任务还引导学生针对特定数据集进行观察，在已有内容和经验上，代入式提供了新的知识和模型思路，并在此基础上探索找到最佳的解决办法，从而引导同学从计算任务中学习更多的知识点和技术方法。

**1.2 任务数据:ScholarSpace中英文学者论文发表数据集**

实验所采用的的数据集收集自ScholarSpace[[1]](#footnote-1)，数据集整体占500MB物理空间存储，共记录102W条论文发表数据。其中中文论文发表情况179382条，记录论文标题、论文合作者、研究领域与发文时间等关键属性，英文论文发表情况846775条，记录论文标题、论文合作者、发文时间、研究领域、论文等级、发文类型等属性。除此之外，数据集还针对每位作者是否出现在论文中文/英文作者列制作了mappingtab和authortab来辅助判断。最后要求针对未对齐的作者信息表，对齐中英文论文中相同的作者信息。

**1.3 阅读文献：**

1. Lin Y, Wang H, Li J, et al. Efficient entity resolution on heterogeneous records[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 32(5): 912-926.
2. Brunner U, Stockinger K. Entity matching on unstructured data: an active learning approach[C]//2019 6th Swiss Conference on Data Science (SDS). IEEE, 2019: 97-102.
3. Köpcke H, Rahm E. Frameworks for entity matching: A comparison[J]. Data & Knowledge Engineering, 2010, 69(2): 197-210.
4. Rastogi V, Dalvi N, Garofalakis M. Large-scale collective entity matching[J]. arXiv preprint arXiv:1103.2410, 2011.
5. Feng D H, Lv Y J, Zhou M. A new approach for English-Chinese named entity alignment[C]. EMNLP2004, 2004.
6. Sun Z, Hu W, Zhang Q, et al. Bootstrapping Entity Alignment with Knowledge Graph Embedding[C]//IJCAI. 2018, 18: 4396-4402.
7. Chen M, Tian Y, Chang K W, et al. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment[J]. arXiv preprint arXiv:1806.06478, 2018.
8. Sun Z, Hu W, Li C. Cross-lingual entity alignment via joint attribute-preserving embedding[C] //International Semantic Web Conference. Springer, Cham, 2017: 628-644.
9. Cao Y, Liu Z, Li C, et al. Multi-channel graph neural network for entity alignment[J]. arXiv preprint arXiv:1908.09898, 2019.
10. Gokhale C, Das S, Doan A H, et al. Corleone: Hands-off crowdsourcing for entity matching[C]//Proceedings of the 2014 ACM SIGMOD international conference on Management of data. 2014: 601-612.
11. Mudgal S, Li H, Rekatsinas T, et al. Deep learning for entity matching: A design space exploration[C]//Proceedings of the 2018 International Conference on Management of Data. 2018: 19-34.
12. Primpeli A, Bizer C, Keuper M. Unsupervised bootstrapping of active learning for entity resolution[C]//European Semantic Web Conference. Springer, Cham, 2020: 215-231.
13. Li H, Konda P, GC P S, et al. MatchCatcher: A Debugger for Blocking in Entity Matching[C]//EDBT. 2018: 193-204.
14. Doan A H, Ardalan A, Ballard J, et al. Human-in-the-loop challenges for entity matching: A midterm report[C]//Proceedings of the 2nd workshop on human-in-the-loop data analytics. 2017: 1-6.
15. Wang J, Li G, Yu J X, et al. Entity matching: How similar is similar[J]. Proceedings of the VLDB Endowment, 2011, 4(10): 622-633.
16. Das S, GC P S, Doan A H, et al. Falcon: Scaling up hands-off crowdsourced entity matching to build cloud services[C]//Proceedings of the 2017 ACM International Conference on Management of Data. 2017: 1431-1446.
17. Shen W, Li X, Doan A H. Constraint-based entity matching[C]//AAAI. 2005: 862-867.
18. Bellare K, Iyengar S, Parameswaran A G, et al. Active sampling for entity matching[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 2012: 1131-1139.
19. Yang K, Liu S, Zhao J, et al. COTSAE: co-training of structure and attribute embeddings for entity alignment[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(03): 3025-3032.
20. Zhao C, He Y. Auto-em: End-to-end fuzzy entity-matching using pre-trained deep models and transfer learning[C]//The World Wide Web Conference. 2019: 2413-2424.

**2. 数据学习（彭迎涛）**

数据学习旨在从大量数据中学习并挖掘深层的语义与潜在价值，强调对数据的分析解读能力。不局限于传统的数据融合，数据学习的场景还包括：（1）方面观点挖掘；（2）细粒度情感分析；（3）语义表征等。本课程中我们设计了开放性计算任务，引导学生进入课程，启发同学通过经典机器学习基本方法、深度学习进阶方法认识任务、实现目标，并最终提升对数据的分析理解能力。

**2.1 计算任务：细粒度情感分析**

情感分析是自然语言处理领域的一项基础任务，具有简单清晰，容易入手的特点。因此，为了快速引导学生进入课程并动手数据学习，我们将细粒度情感分析作为数据学习模块的典型计算任务，以提升学生数据的分析、理解和处理的能力。

在线评论的细粒度情感分析对于深刻理解商家和用户、挖掘用户情感等方面有至关重要的价值，并在互联网平台主营业务中广泛应用，包括个性化推荐、智能搜索、产品反馈等。由此，细粒度情感分析的计算任务，我们从以下方面引入并开展：

（1）任务目标：根据已标注细粒度要素的情感倾向建立模型算法，对用户评论实现细粒度情感挖掘。

（2）基础模型：本次任务提供了机器学习方法的基础分类模型（SVM），帮助同学对任务认识和理解。

（3）结果测评：本次任务提供了测试平台，并提供了基准代码（baseline），评价结果为：0.36。在此基础上进行改进并提交结果，系统将所有学生提交结果给出排名（展示基准代码的代码片段、排名截图）。目前课程中已经获得的最高分为：0.59。

数据学习部分代码

|  |
| --- |
| Baseline |
| for col in tqdm(columns[1:]):  print(col)  train\_label=train[col]  valid\_label=valid[col]  mymodel=model(train\_features,train\_label,modelname="SVM")  mymodel.train()  preds=mymodel.predict(valid\_features)  ff=f1\_score(valid\_label, preds, average='macro')  print('f1=%s'%ff)  f1+=ff  d\_test[col]=mymodel.predict(test\_features)  print('平均 f1值=%s'%(f1/10))    class model():      def \_\_init\_\_(self,train\_features , train\_label,modelname='SVM'):          self.train\_features=train\_features          self.train\_label=train\_label          self.modelname=modelname          self.mymodel=None          if modelname=='SVM':              self.mymodel=SVC(kernel= "rbf")          elif modelname=='randomforest':              self.mymodel=RandomForestRegressor()      def train(self):          print('开始训练...')          self.mymodel.fit(self.train\_features,self.train\_label)      def predict(self,valid\_features):          preds=self.mymodel.predict(valid\_features)          preds=[np.round(lb) for lb in preds ]          return preds |



数据学习计算任务测试平台页面

（4）友情提示：

1.选用词嵌入分类模型（如：Fasttext），可提升0.06左右效果；2.采用适用的机器学习模型（如：XGBoost），可提升0.12左右；3.选用词嵌入结合深度学习模型（如：Word2vec+GRU），可提升0.16左右；4选用预训练模型，可提升0.2左右。

为进一步提升学生的数据分析处理能力，本任务还引导学生针对特定数据集进行观察，在已有内容和经验上，代入式提供了新的知识和模型思路，并在此基础上探索找到最佳的解决办法，从而引导同学从计算任务中学习更多的知识点和技术方法。

**2.2 任务数据：美团点评的评论数据**

数据集分为训练、验证、测试三个部分。数据集中的评价对象按照粒度不同划分成两个层次，层次一为粗粒度的评价对象，例如评论文本中涉及的服务、位置等要素；层次二为细粒度的情感对象，例如“服务”属性中的“服务人员态度”、“排队等候时间”等细粒度要素。每个细粒度要素的情感倾向有四种状态：正向(Positive)、中性(Neutral)、负向(Negative)、未提及(Not mentioned)，使用[1, 0, -1, -2]四个值对情感倾向进行表示。详情见下表 1 美团评论细粒度要素表[[2]](#footnote-2)。

表 1 美团评论细粒度要素表

|  |  |
| --- | --- |
| level 1 | level 2 |
| location | location\_traffic\_convenience |
| location\_distance\_from\_business\_district |
| location\_easy\_to\_find |
| service | service\_wait\_time |
| service\_waiters\_attitude |
| service\_parking\_convenience |
| service\_serving\_speed |
| price | price\_level |
| price\_cost\_effective |
| price\_discount |
| environment | environment\_decoration |
| environment\_noise |
| environment\_space |
| environment\_cleaness |
| dish | dish\_portion |
| dish\_taste |
| dish\_look |
| dish\_recommendation |
| other | others\_overall\_experience |
| others\_willing\_to\_consume\_again |

**2.3 阅读文献:**

1. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
2. Le Q, Mikolov T. Distributed representations of sentences and documents[C]//International conference on machine learning. PMLR, 2014: 1188-1196.
3. Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.
4. Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification[C]//Twenty-ninth AAAI conference on artificial intelligence. 2015.
5. Tang D, Qin B, Feng X, et al. Effective LSTMs for target-dependent sentiment classification[J]. arXiv preprint arXiv:1512.01100, 2015.
6. Zhang D, Xu H, Su Z, et al. Chinese comments sentiment classification based on word2vec and SVMperf[J]. Expert Systems with Applications, 2015, 42(4): 1857-1863.
7. Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification[J]. Advances in neural information processing systems, 2015, 28.
8. Zhou C, Sun C, Liu Z, et al. A C-LSTM neural network for text classification[J]. arXiv preprint arXiv:1511.08630, 2015.
9. Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning[J]. arXiv preprint arXiv:1605.05101, 2016.
10. Tang D, Qin B, Liu T. Aspect level sentiment classification with deep memory network[J]. arXiv preprint arXiv:1605.08900, 2016.
11. Wang J, Yu L C, Lai K R, et al. Dimensional sentiment analysis using a regional CNN-LSTM model[C]//Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers). 2016: 225-230.
12. Wang Y, Huang M, Zhu X, et al. Attention-based LSTM for aspect-level sentiment classification[C]//Proceedings of the 2016 conference on empirical methods in natural language processing. 2016: 606-615.
13. Conneau A, Schwenk H, Barrault L, et al. Very deep convolutional networks for text classification[J]. arXiv preprint arXiv:1606.01781, 2016.
14. Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 562-570.
15. Nowak J, Taspinar A, Scherer R. LSTM recurrent neural networks for short text and sentiment classification[C]//International Conference on Artificial Intelligence and Soft Computing. Springer, Cham, 2017: 553-562.
16. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
17. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
18. Howard J, Ruder S. Universal language model fine-tuning for text classification[J]. arXiv preprint arXiv:1801.06146, 2018.
19. Liu Q, Huang H Y, Gao Y, et al. Task-oriented word embedding for text classification[C]//Proceedings of the 27th international conference on computational linguistics. 2018: 2023-2032.
20. Wang B, Lu W. Learning latent opinions for aspect-level sentiment classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).
21. Du C, Sun H, Wang J, et al. Capsule network with interactive attention for aspect-level sentiment classification[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 5489-5498.
22. Li X, Bing L, Zhang W, et al. Exploiting BERT for end-to-end aspect-based sentiment analysis[J]. arXiv preprint arXiv:1910.00883, 2019.
23. Sun C, Huang L, Qiu X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence[J]. arXiv preprint arXiv:1903.09588, 2019.
24. Sun K, Zhang R, Mensah S, et al. Aspect-level sentiment analysis via convolution over dependency tree[C]//Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). 2019: 5679-5688.

**3. 数据增强（王文礼）**

数据增强作为一种数据扩充技术，旨在让有限的数据产生更多的等价数据，是高质量数据融合与知识融合的基础之一，其主要应用场景可以归纳为：(1)大样本数据需求场景；（2）高质量数据需求场景。为了提升模型效果，众多研究者在模型改进中提出了诸多优秀方法，但较低质量的数据将限制模型效果，数据增强的重要性不言而喻。据此，本课程依托开放性研究场景，进而引导学生进入课程，启发学生如何通过合理地修正数据、扩充数据等方法，完成数据增强，最终达到提升学生自主学习、研究能力。

**3.1 计算任务：小样本数据增强**

根据数据增强应用场景，从实际需求出发，该部分设置的课程任务为小样本数据增强。为了引导同学们快速进入课程，在该任务中，我们为同学们提供大量的有噪声数据（标签被扰动）以及少量的无噪声数据（Ground truth），启发同学带入并展开数据增强任务的探索。具体内容如下。

（1）任务目标：对少量无噪声数据样本进行数据增强，从而用于清洗、修正有噪声数据的标签。旨在引导学生通过数据增强方法优化训练数据集的质量，提升该数据在黑盒模型上的评测效果——准确率。

（2）基础模型：本例为开放性研究案例，同学们可从数据到模型训练完整过程中的任一环节出发，提出自己的模型进而完成优化数据，也可以将问题转化为无监督学习问题，进而优化；同时也可从数据本身出发，直接对数据优化。

（3）效果测评：课程为本次任务提供黑盒测试平台，学生仅需将优化后的数据提交平台，平台使用提交数据训练模型并返回预测精度，并展示测评效果和排名。此外，课程为学生提供基准代码（baseline），预测结果在0.68左右。

（4）课程提示：①数据角度：直接对给出的Ground truth进行数据增强，该方法约可提升精度0.05。②方法角度：利用给出的Ground truth训练模型，对噪声数据清洗，约可提升精度0.10。③数据角度：将问题转化为无监督数据标注问题，可提升0.03。④：混合方法：多种方法可混合使用。

课程该部分内容将应用中的需求抽象为计算任务，以此激发学生探索热情，帮助学生快速高效掌握数据增强的思想，为后续学习与科研打下基础。

**3.2 任务数据：中文新闻数据**

数据集选用已标注的中文新闻数据[[3]](#footnote-3)，涵盖教育、科技、时政、财经、家居、娱乐、房产、星座、彩票、时尚、游戏、体育、社会、股票共14领域。我们对原始数据集进行裁剪，得到12，000条数据样本，使用数据标签0-13表示。数据集共包含为两部分，一类是无噪声的数据2000条，一类是标签随机打乱的数据约12000条。数据集的使用存在两大难点：首先，可用的无噪声数据较少，训练高精度模型相对困难；其次，中文蕴含的丰富语义使得可借鉴模型相对较少，数据利用难度大。

**3.3 阅读文献：**

1. Anaby-Tavor A, Carmeli B, Goldbraich E, et al. Do not have enough data? Deep learning to the rescue![C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(05): 7383-7390.
2. De Angeli K, Gao S, Danciu I, et al. Class imbalance in out-of-distribution datasets: Improving the robustness of the TextCNN for the classification of rare cancer types[J]. Journal of biomedical informatics, 2022, 125: 103957.
3. Chen J, Yang Z, Yang D. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification[J]. arXiv preprint arXiv:2004.12239, 2020.
4. Feng S Y, Gangal V, Kang D, et al. GenAug: Data Augmentation for Finetuning Text Generators[C]//Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. 2020: 29-42.
5. Guo H. Nonlinear mixup: Out-of-manifold data augmentation for text classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(04): 4044-4051.
6. Gururangan S, Dang T, Card D, et al. Variational Pretraining for Semi-supervised Text Classification[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 5880-5894.
7. Karimi A, Rossi L, Prati A. AEDA: An Easier Data Augmentation Technique for Text Classification[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. 2021: 2748-2754.
8. Kedzie C, Mckeown K. A Good Sample is Hard to Find: Noise Injection Sampling and Self-Training for Neural Language Generation Models[C]//Proceedings of the 12th International Conference on Natural Language Generation. 2019: 584-593.
9. Kim D, Koo J, Kim U M. EnvBERT: Multi-Label Text Classification for Imbalanced, Noisy Environmental News Data[C]//2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM). IEEE, 2021: 1-8.
10. Kumar V, Choudhary A, Cho E. Data Augmentation using Pre-trained Transformer Models[C]//Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems. 2020: 18-26.
11. Li Y, Cohn T, Baldwin T. Robust training under linguistic adversity[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. 2017: 21-27.
12. Liu R, Xu G, Jia C, et al. Data Boost: Text Data Augmentation Through Reinforcement Learning Guided Conditional Generation[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 9031-9041.
13. Montella S, Fabre B, Urvoy T, et al. Denoising Pre-Training and Data Augmentation Strategies for Enhanced RDF Verbalization with Transformers[C]//Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+). 2020: 89-99.
14. Olsson V, Tranheden W, Pinto J, et al. Classmix: Segmentation-based data augmentation for semi-supervised learning[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 1369-1378.
15. Padurariu C, Breaban M E. Dealing with data imbalance in text classification[J]. Procedia Computer Science, 2019, 159: 736-745.
16. Quteineh H, Samothrakis S, Sutcliffe R. Textual data augmentation for efficient active learning on tiny datasets[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020: 7400-7410.
17. Ren S, Deng Y, He K, et al. Generating natural language adversarial examples through probability weighted word saliency[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. 2019: 1085-1097.
18. Ren S, Zhang J, Li L, et al. Text AutoAugment: Learning Compositional Augmentation Policy for Text Classification[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 9029-9043.
19. Wei J, Zou K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks[J]. arXiv preprint arXiv:1901.11196, 2019.
20. Zhang R, Yu Y, Zhang C. SeqMix: Augmenting Active Sequence Labeling via Sequence Mixup[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 8566-8579.

**4. 数据实时分析（张旭康）**

数据分析技术旨在从分析数据的过程和结果中获得有用的信息，用于让分析者产生认识和做出决策。数据分析按照时间特点可以分为实时数据分析和离线数据挖掘。在实时数据分析中，流数据的分析为其中一个大类。如在科学研究、智慧交通、无人驾驶、城市监控等领域，大规模分布式的传感器、监控摄像等装置的建设，产生了高速、海量的流数据。从流数据中进行实时分析，发现热点、稀有事件，依靠传统的人工+计算机搜索已无能为力。其中实时场景下的目标发现也成为了实时流数据处理中的重点和热点。近年来发展并盛行起来的数据智能技术正是从大数据中解决这一问题的利器。离线数据挖掘则是对大量累积的数据进行深入的数据分析，使用更复杂的技术，从数据中寻找更有价值的目标。

**4.1 计算任务：科学数据智能发现**

本案例提供由70万余条天文数据（光变曲线）。分为实时和离线数据分析两种场景。在实时场景下，提供一组异常曲线，学生需要设计算法以在异常光变曲线生成的过程中检测到正在发生的科学现象。在离线场景下，学生需要使用智能分析、数据挖掘的方法，从海量数据中尽可能挖掘稀少的科学现象。科学现象主要包括微引力透镜候选体和恒星耀发候选体这两种短时标稀有天体光变事件，计算任务为正确标识稀有现象发生的范围，以及识别出包含以上两种稀有天体光变事件（即一段异常子序列）的时间序列。

（1）任务目标：分别在实时和离线场景下从天文数据中标注异常序列和挖掘异常时间序列。

（2）基础模型：本次任务提供了实时和离线场景下的baseline。

（3）结果测评：本次任务提供了测试平台，学生可将计算结果提交平台，查看测评效果和排名。实时场景下，事先标记异常段，然后根据异常段的检测准确度打分。离线场景下，根据检测出的稀有曲线数量和真实度的比例打分。

 利用光变曲线稀有事件探测，激发学生处理数据的兴趣，启发学生通过处理数据作为“眼睛”来看到更多有用的信息，探索更广阔的科学世界。从而培养学生的数据素养。

**4.2 任务数据：GWAC光变曲线数据**

本案例的数据来源于中国科学院国家天文台自主研发的地基广角相机阵(GWAC，Ground- based Wide Angle Camera)，已公布于阿里天池实验室[[4]](#footnote-4)。该设备每15秒钟采样一份观测数据，并得益于其大视场覆盖能力，至今已获得具有15秒采样分辨率的数百万条光变样本。本案例提供的数据属性为包含有时间信息和天体目标光度信息的时域数据(光变曲线)。

数据挑战：1、由于是真实观测数据，不可避免地会有部分数据含有不可去除的噪声，这部分噪声数据也可能会对识别造成干扰，如光变曲线短暂的下降等；2、由于真实观测受环境影响很大，如白天或大量云雾遮挡，导致光变曲线也有可能是不连续的，即相邻两个观测数据点的时间跨度非常大，中间会缺失了很多数据点。

微引力透镜候选体表现为天体光度逐步变亮而后逐步恢复正常的过程，且变化部分对称，如图1所示。恒星耀发候选体表现为天体快速变亮而后缓慢恢复的过程，且变化部分是非对称的，如图2所示。

|  |  |
| --- | --- |
| 图1 微引力透镜的光变曲线形状 | 图2 恒星耀发的光变曲线形状 |

**4.3 阅读文献：**

1. Qiu J, Sun Y, Wu C, et al. NFD: Toward real-time mining of short-timescale gravitational microlensing events[J]. Publications of the Astronomical Society of the Pacific, 2018, 130(992): 104504.
2. 杨晨,翁祖建,孟小峰,任玮,忻日辉,王春凯,都志辉,万萌,魏建彦.天文大数据挑战与实时处理技术[J].计算机研究与发展,2017,54(02):248-257.
3. Davenport J , Hawley S L , Hebb L , et al. Kepler Flares II: TheTemporal Morphology of White-Light Flares on GJ 1243. TheAstrophysical Journal, 2014, 797(2):122
4. Richards J W, Starr D L, Butler N R, et al. On machine-learned classification of variable stars with sparse and noisy time-series data[J]. The Astrophysical Journal, 2011, 733(1): 10.
5. 梁恩思.恒星耀发的测光观测及系外行星的证认[D].南京大学, 2020.
6. 孟小峰.科学数据智能：人工智能在科学发现中的机遇与挑战[J].中国科学基金,2021,35(03):419-425.DOI:10.16262/j.cnki.1000-8217.20210611.006.
7. 黎建辉,沈志宏,孟小峰.科学大数据管理:概念、技术与系统[J].计算机研究与发展,2017,54(02):235-247.
8. 陶一寒,崔辰州,张彦霞,许允飞,樊东卫,韩叙,韩军,李长华,何勃亮,李珊珊,米琳莹,杨涵溪,杨丝丝.深度学习在天文学中的应用与改进[J].天文学进展,2020,38(02):168-188.

**5.知识图谱（艾山）**

知识图谱是组织知识的一种形式之一。通过知识图谱构建技术可以把非结构化数据转换成结构化形式从而有效管理、检索和利用非结构化数据。知识图谱被用于多个领域，且已有广泛研究。构建知识图谱主要由（1）实体识别；（2）关系抽取；两个部分组成。本课程中我们设计了开放性计算任务，引导学生进入课程，启发同学通过经典机器学习基本方法、深度学习进阶方法认识任务、实现目标，并最终理解基于知识图谱的数据组织能力。

**5.1 计算任务：微生物领域关系抽取**

在知识图谱构建任务中关系抽取比较重要的，也就是关系发现，通过关系发现我们可以体会到实体之间的语义关系，不断的扩展知识图谱，从而为下游任务提供知识，比如：推荐系统，问答系统，搜索引擎等。由此，本课程引入微生物领域关系抽取计算任务，我们从以下方面引入并开展：

（1）任务目标：根据给定的实体对和pubmed文本语料标注数据并建立模型算法，从文本语料中发现更多的关系。

（2）基础模型：本次任务提供了机器学习方法的基础分类模型（SVM），帮助同学对任务认识和理解。

（3）结果测评：本次任务提供了测评平台，同学们可将自己的模型计算结果提交到平台上，并查看测评分数和排名。

为进一步提升学生的数据分析处理能力，本课程从任务出发，在已有内容和经验上，代入式提供了新的知识和模型思路，引导同学从计算任务中学习更多的知识点和技术方法。

**5.2 任务数据：微生物实体对+文本语料**

我们的提供66万微生物实体对（细菌，栖息地），以及大量的文本语料（Pubmed）和公开测试数据。同学们实体对和文本语料远程监督地标注好数据并且将数据集分为训练、验证、测试三个部分。公开数据是以句子形式，每一条句子包含一对实体（细菌，栖息地）和上下文，并有关系标签。

实体对数据来自微生物所所提供；文本数据由wamdm实验室收集公开的学术文献摘要。

**5.3 阅读文献：**

1. Augenstein, I., Maynard, D., & Ciravegna, F. (2016). Distantly supervised Web relation extraction for knowledge base population. SEMANTIC WEB, 7(4), 335–349. <https://doi.org/10.3233/SW-150180>
2. Bakal, G., Talari, P., Kakani, E., V., & Kavuluru, R. (2018). Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations. JOURNAL OF BIOMEDICAL INFORMATICS, 82, 189–199. <https://doi.org/10.1016/j.jbi.2018.05.003>
3. Dumitrache, A., Aroyo, L., & Welty, C. (2018). Crowdsourcing Ground Truth for Medical Relation Extraction. ACM TRANSACTIONS ON INTERACTIVE INTELLIGENT SYSTEMS, 8(2). <https://doi.org/10.1145/3152889>
4. Fan, M., Zhao, D., Zhou, Q., Liu, Z., Zheng, T. F., & Chang, E. Y. (2014). Distant Supervision for Relation Extraction with Matrix Completion (K. Toutanova & H. Wu, Eds.; WOS:000493814100079; pp. 839–849).
5. Feng, X., Guo, J., Qin, B., Liu, T., & Liu, Y. (2017). Effective Deep Memory Networks for Distant Supervised Relation Extraction (C. Sierra, Ed.; WOS:000764137504018; pp. 4002–4008).
6. Han, X., Sun, L., & AAAI. (2016). Global Distant Supervision for Relation Extraction (WOS:000485474202138). 2950–2956.
7. He, Z., Chen, W., Li, Z., Zhang, M., Zhang, W., Zhang, M., & AAAI. (2018). SEE: Syntax-Aware Entity Embedding for Neural Relation Extraction (WOS:000485488905110). 5795–5802.
8. Lamurias, A., Clarke, L. A., & Couto, F. M. (2017). Extracting microRNA-gene relations from biomedical literature using distant supervision. PLOS ONE, 12(3). <https://doi.org/10.1371/journal.pone.0171929>
9. Li, Y., Long, G., Shen, T., Zhou, T., Yao, L., Huo, H., Jiang, J., & Assoc Advancement Artificial Intelligence. (2020). Self-Attention Enhanced Selective Gate with Entity-Aware Embedding for Distantly Supervised Relation Extraction (WOS:000668126800079). 34, 8269–8276.
10. Lin, Y., Shen, S., Liu, Z., Luan, H., & Sun, M. (2016). Neural Relation Extraction with Selective Attention over Instances (K. Erk & N. Smith, Eds.; WOS:000493806800200; pp. 2124–2133).
11. Lockard, C., Dong, X. L., Einolghozati, A., & Shiralkar, P. (2018). CERES: Distantly Supervised Relation Extraction from the Semi-Structured Web. PROCEEDINGS OF THE VLDB ENDOWMENT, 11(10), 1084–1096. <https://doi.org/10.14778/3231751.3231758>
12. Luo, B., Feng, Y., Wang, Z., Zhu, Z., Huang, S., Yan, R., & Zhao, D. (2017). Learning with Noise: Enhance Distantly Supervised Relation Extraction with Dynamic Transition Matrix (R. Barzilay & M. Kan, Eds.; WOS:000493984800040; pp. 430–439). <https://doi.org/10.18653/v1/P17-1040>
13. Qin, P., Xu, W., & Wang, W. Y. (2018). Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning (I. Gurevych & Y. Miyao, Eds.; WOS:000493904300199; pp. 2137–2147).
14. Qu, J., Ouyang, D., Hua, W., Ye, Y., & Li, X. (2018). Distant supervision for neural relation extraction integrated with word attention and property features. NEURAL NETWORKS, 100, 59–69. <https://doi.org/10.1016/j.neunet.2018.01.006>
15. Quirk, C., Poon, H., & Assoc Computat Linguist. (2017). Distant Supervision for Relation Extraction beyond the Sentence Boundary (WOS:000712360800110). 1171–1182.
16. Wu, S., Fan, K., Zhang, Q., & AAAI. (2019). Improving Distantly Supervised Relation Extraction with Neural Noise Converter and Conditional Optimal Selector (WOS:000486572501100). 7273–7280.
17. Xie, Y., Xu, H., Li, J., Yang, C., & Gao, K. (2020). Heterogeneous graph neural networks for noisy few-shot relation classification. KNOWLEDGE-BASED SYSTEMS, 194. <https://doi.org/10.1016/j.knosys.2020.105548>
18. Yaghoobzadeh, Y., Adel, H., Schuetze, H., & Assoc Computat Linguist. (2017). Noise Mitigation for Neural Entity Typing and Relation Extraction (WOS:000712360800111). 1183–1194.
19. Ye, H., & Luo, Z. (2020). Deep ranking based cost-sensitive multi-label learning for distant supervision relation extraction. INFORMATION PROCESSING & MANAGEMENT, 57(6). <https://doi.org/10.1016/j.ipm.2019.102096>
20. Yuan, Y., Liu, L., Tang, S., Zhang, Z., Zhuang, Y., Pu, S., Wu, F., Ren, X., & AAAI. (2019). Cross-Relation Cross-Bag Attention for Distantly-Supervised Relation Extraction (WOS:000485292600052). 419–426.
21. Zeng, D., Dai, Y., Li, F., Sherratt, R. S., & Wang, J. (2018). Adversarial Learning for Distant Supervised Relation Extraction. CMC-COMPUTERS MATERIALS & CONTINUA, 55(1), 121–136. <https://doi.org/10.3970/cmc.2018.055.121>
22. Zeng, X., He, S., Liu, K., Zhao, J., & AAAI. (2018). Large Scaled Relation Extraction with Reinforcement Learning (WOS:000485488905093). 5658–5665.
23. Zhang, X., Liu, H., & Wu, Z. (2020). Noise Reduction Framework for Distantly Supervised Relation Extraction with Human in the Loop (L. Wenzheng & Z. Xuefei, Eds.; WOS:000577101800001; pp. 1–4). <https://doi.org/10.1109/ICEIEC49280.2020.9152287>
24. Zhou, P., Xu, J., Qi, Z., Bao, H., Chen, Z., & Xu, B. (2018). Distant supervision for relation extraction with hierarchical selective attention. NEURAL NETWORKS, 108, 240–247. <https://doi.org/10.1016/j.neunet.2018.08.016>

1. http://cdblp.ruc.edu.cn/ [↑](#footnote-ref-1)
2. https://challenger.ai/ [↑](#footnote-ref-2)
3. 中文文本分类数据集THUCNews：http://thuctc.thunlp.org [↑](#footnote-ref-3)
4. https://tianchi.aliyun.com/competition/entrance/531805/introduction [↑](#footnote-ref-4)