**临床数据集偏移文献综述**

【摘要】随着医疗大数据时代的来临，越来越多的高质量临床数据集由专业机构整理收集，形成了一套完整的数据库，这些数据库促进了临床资源的使用，为了人类医疗事业做出了重大的贡献。然而在使用这些医疗数据库进行建模和科研的过程中，科学家们也经常遇到使用公开医疗数据库进行训练的模型在实际应用于测试数据集时结果很差的情况，即数据集偏移的问题。造成这种临床数据集偏移的原因可能是训练样本与测试样本在变量空间的分布差异巨大，即特征值偏移，或训练数据集与测试数据集的类别分布不均衡，即类别偏移。通过对临床数据集基于性别、年龄、社会经济状况等客观标准进行患者分层并与测试数据集进行比对，进而筛选合适的临床数据集或调整不同患者分层的占比可以在一定程度上降低临床数据集偏移的影响。

【关键词】临床数据集、数据集偏移、临床数据集偏移

【Abstract】With the advent of the era of medical big data, more and more high-quality clinical datasets are organized and collected by professional institutions to form a complete set of databases. These datasets promote the use of clinical resources and make significant contributions to the human medical industry. However, in the process of modeling and research using these clinical datasets, scientists often encounter the problem of Dataset Shift, that is when models trained using public clinical datasets yield unsatisfactory results when applied to test datasets. The reason for Clinical Dataset Shift may be due to significant differences in the distribution of training and test datasets in the variance space, namely Co-variate Shift or imbalanced class distribution between the training and test datasets, namely Prior Probability Shift. By stratifying patients based on objective criteria such as gender, age and socio-economic status in clinical datasets and comparing them with test datasets, thereby selecting suitable clinical datasets or adjusting the proportion of different patient stratification can to some extent reduce the effect of Clinical Dataset Shift.

【Keywords】Clinical Datasets、Dataset Shift、Clinical Dataset Shift

【综述提纲】

1. 临床数据集应用现状
2. 临床数据集偏移对模型训练、结果诊断的不良影响
3. 临床数据集偏移的产生原因分析
4. 消除或缓解临床数据集偏移的方法