# Dynamic Semantic Annotation Method for Sensor Data Integrating PCA and SVM*

ZHANG Shuyue, DUAN Chao*
(School of Engineering, Guangzhou College of Technology and Business, Foshan Guangdong 528100, China)

**Abstract**: The long-range and short-range dependencies between data elements collected by sensing devices are variable, and their dynamism makes it difficult to track the data annotation process in real time, limiting accuracy. To this end, a dynamic semantic annotation method for sensor data is proposed, which integrates principal component analysis (PCA) and support vector machine (SVM). Sensor data are integrated and optimized, including data fluctuation suppression, intelligent reconstruction of missing data, and scale unification. PCA algorithm is used to perform dimensionality reduction on the optimized data, reducing the number of features while preserving key information. A semi supervised active learning method based on adaptive SVM, combining labeled and unlabeled data, continuously optimizes classifier performance through active learning and semi supervised learning, adapts to the dynamic nature of sensor data, and achieves dynamic semantic annotation of sensor data. The experimental results show that the proposed method has a high macro mean of 36.8 under all sample percentage conditions, and can accurately output semantic annotation results.

**Key words**: data collection and processing; PCA dimensionality reduction; ASVM-SAL algorithm; dynamic semantic annotation
**EEACC**: 7230 **doi**: 10.3969/j.issn.1004-1699.2025.11.023

# 融合 PCA 与 SVM 的传感数据动态语义标注方法*

张书月, 段 超*
(广州工商学院工学院, 广东 佛山 528100)

**摘 要**: 传感设备采集数据元素间的长距离依赖联系和短距离依赖联系多变, 其动态性使得数据标注过程难以实时跟踪, 准确性受限。为此, 提出融合主成分分析(Principal Component Analysis, PCA)与支持向量机(Support Vector Machine, SVM)的传感数据动态语义标注方法。对传感数据进行集成与优化, 包括数据波动抑制、缺失数据智能重构和尺度统一化。采用 PCA 算法对优化后的数据进行降维处理, 减少特征数量并保留关键信息。基于自适应 SVM 的半监督主动学习方法, 结合标注和未标注数据, 通过主动学习和半监督学习不断优化分类器性能, 适应传感数据动态性, 实现对传感数据的动态语义标注。实验结果表明, 所提方法在所有样本百分比条件下均具有较高的宏平均值, 平均为 36.8, 可精准输出语义标注结果。

**关键词**: 数据采集和处理; PCA 降维; ASVM-SAL 算法; 动态语义标注
**中图分类号**: TP391 **文献标识码**: A **文章编号**: 1004-1699(2025)11-2094-06

传感器设备能够实时捕获大量、连续且异构的数据流, 为各种应用提供了丰富的数据源[1]。由于传感器设备不断采集数据, 并且数据信息实时变化, 因此传感数据呈现出高度的动态性。这种动态性使得传统的静态数据标注方法无法有效地满足需求。静态数据标注通常依赖于人工操作, 虽然能够实现对数据的精准标注, 但面对海量、连续的传感数据流, 人工标注不仅耗时费力, 而且难以及时满足数据标注的需求。因此, 如何从这些海量数据中提取出有价值的信息, 并根据上下文进行准确、高效的语义解释, 是当前物联网领域面临的重要挑战之一。

胡青宁等[2]采用关键字驱动的传感数据语义描述方式, 抽取目标词汇、属性词汇等构成数据关键字集, 以此集合来阐释数据含义; 结合编辑距离与关键字项集相似度计算数据元语义相似度; 构建数据元字典, 实现传感数据到数据模型的动态语义标注。但该方法编辑距离仅考虑字符层面的差异, 无法捕捉到词汇间更复杂的语义动态关系, 准确性不足。费豪等[3]运用动态句法精简策略, 利用递归神经网络或具备句法标注的图卷积网络对传感数据的句法构造实施编

**项目来源**: 2024 广州工商学院校级科研重点项目(KYZD202403); 2024 年佛山市自筹经费类科技创新项目(2420001004716); 广东省重点建设学科科研能力提升项目(2024ZDJS087); 十五运会和残特奥会粤港澳协同发展理论与实践研究课题(2025GBA-220)
**收稿日期**: 2025-03-14 **修改日期**: 2025-05-15

码;通过离散化句法连接,实现动态句法裁剪;将裁剪后的句法特征融入标注模型,完成传感数据的动态语义标注。该方法将句法特征融入标注模型,复杂的特征选择与融合过程难以适应传感数据动态性,增加了算法实现的难度,降低了准确性。Xu 等[4]通过改进粗糙概念格构建传感数据语义本体;通过本体映射集成模块集成本体,信息提取模块提取信息,实现传感数据的动态语义标注。但该方法在解析数据过程中耗时较长,难以满足快速变化的标注需求。Xu 等[5]构建异构信息传感网络整合服务、提供者与用户关联,提出 GoT 模型。该模型通过元路径构建整合数据,元路径内与元路径间融合提取特征,实施语义注释推荐。但该方法元路径内与元路径间的特征提取受限于传感数据,如果数据快速变化,单一特征不够可靠,进而影响语义标注的效果。

为此,提出融合 PCA 与 SVM 的传感数据动态语义标注方法。通过 PCA 算法对传感数据进行降维处理,有效减少了特征数量,同时保留了数据的关键信息,为后续的分类任务奠定了坚实基础。降维过程中,利用坡度图确定最佳主成分个数,确保了降维效果的最优化。采用基于自适应 SVM 的半监督主动学习方法,结合标注和未标注数据,通过主动学习和半监督学习的迭代优化,不断提升分类器的性能。该方法能够充分利用未标注数据中的有用信息,通过构建加权矩阵和计算谐波函数等方式,实现对未标注数据的精准标注。在每一轮标注后,根据风险评估结果选择最优标注样本,进一步提高了标注的准确性和效率。输出一个经过优化的自适应分类器及其对应的标注结果。该方法利用自适应 SVM 结合半监督主动学习,能根据数据动态性自适应调整,充分利用标注与未标注数据,解决单一标注偏移问题,实现对传感数据的实时精准标注。

## 1　传感数据集成与优化

在传感网络中,广泛分布于各监控领域的传感器节点具有集成环境参数(如温度、湿度、压力等关键指标)或设备状态情报的作用。图 1 展示无线传感网络的基本构成。

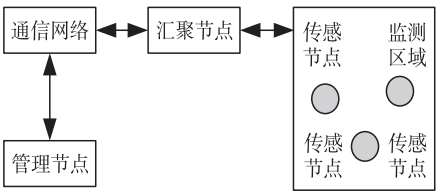


**图 1　无线传感网络组成结构**

为了高效集成这些信息,选取网络中的战略节点或集线节点(如传感器节点、中继站等)作为信息集成的核心[6]。根据信息的生成速率和综合集成需求,配置集成频率 $g$,确保所有必要的传感信息被精准集成。集成频率的配置公式如下:

$$\begin{cases} g=\dfrac{M_w}{M_j} \\ M_j \leqslant M_w \end{cases} \tag{1}$$

式中:$M_w$ 为无线传感网络中实际传感数据量;$M_j$ 为待集成的传感数据量。通过配置截获频率和过滤条件,确保集成所有必要的传感数据。

设定集成点数量 $w$、集成频率 $g$ 和集成时间 $\Delta t$,使用下式计算传感数据的总集成规模 $x$:

$$x=\sum_{i=1}^{w} g\times x_i\times\Delta t \tag{2}$$

式中:$x_i(i=1,2,3,\cdots,w)$ 为第 $i$ 个集成点在单位时间内汇聚的传感数据量。

在信息集成完毕后,需对传感信息实施一系列深度优化,提升其精确度和信赖度。

①数据波动抑制:

采用先进的均值滤波算法对数据展开平滑增强以抑制波动[7],具体表达式如下:

$$x_l=\frac{1}{Q}\sum_{i=1}^{w} x_i \tag{3}$$

式中:$Q$ 为网络运行数据集成量。

②数据缺失智能重构与填补:

采用线性插值技术,对缺失数据展开高精度重构与填补。重构填补公式为:

$$x_b(k)=\frac{x(k-1)+x(k+1)}{2} \tag{4}$$

式中:$k$ 为缺失数据位置;$x(k-1)$,$x(k+1)$ 为 $k$ 前后两个相邻数据。通过智能重构,确保了数据的连续性和完整性。

③数据尺度统一化[8]:

采用高级归一化算法,对数据展开尺度统一处理。具体公式为:

$$x_g=\frac{\max(x)-x}{\max(x)-\min(x)} \tag{5}$$

式中:$\max(x)$,$\min(x)$ 分别为初始集成数据中的最大值和最小值。

经过上述数据集成与深度优化步骤后,获得了优化后的传感数据。将这些数据记为 $x_g$,并作为后续分析算法的主要输入。由此,这些预处理步骤有效消除了数据中的噪声和异常值,确保了数据的完整性和连续性,为后续的降维和分类提供了可靠的基础。

## 2 传感数据动态语义标注方法

### 2.1 基于 PCA 算法的数据降维

经过上节处理后的传感数据 $x_g$,其中包含大量的参数信息,导致数据集包含众多特征或变量,具有高维度特性。PCA 通过线性变换,将原始的高维数据投影到一个新的低维坐标系中,使得新坐标系下的各维度(即主成分)之间相互正交,并且按照数据方差的大小依次排列。在传感数据的研究中,传感器会采集大量的参数信息,导致数据集具有高维度特性。高维数据不仅增加了数据存储和处理的成本,还可能带来“维度灾难”问题,即随着维度的增加,数据在空间中的分布变得稀疏。PCA 技术能够有效应对这些问题,将数据从高维空间转换到低维空间,同时尽可能保留数据的主要特征。通过降低数据维度,减少了后续分类器需要处理的特征数量,降低了计算复杂度,提高了算法的运行效率。

为此,采用 PCA 技术处理上节获得的传感数据,缩减数据维度[9-10],减少了特征数量,同时保留了关键信息。由此简化了数据模型,降低了后续分类器的计算复杂度,同时避免了过拟合的风险。

基于上节处理后的传感数据集合 $x_g$,包含 $n$ 个样本和 $m$ 个传感指标数据。将数据保存为矩阵:

$$\boldsymbol{X}=\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1m} \\ X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nm} \end{bmatrix} \tag{6}$$

式中:$X_{ij}$为第 $i(i=1,2,3,\cdots,n)$ 个样本的第 $j(j=1,2,3,\cdots,m)$ 个传感指标值。计算每列(即每个传感指标)的平均值 $M_j$ 和样本数据值方差 $B=\left[\sum_{i=1}^{n}(X_{ij-A})^2/(n-1)\right]$,形成一个 $m$ 维矩阵 $\boldsymbol{M}$,其中每个元素代表对应的传感指标值:

$$\boldsymbol{M}=\begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_m \end{bmatrix}=\begin{bmatrix} n^{-1}\sum_{i=1}^{n}X_{i1} \\ n^{-1}\sum_{i=1}^{n}X_{i2} \\ \vdots \\ n^{-1}\sum_{i=1}^{n}X_{\mathrm{im}} \end{bmatrix} \tag{7}$$

PCA 通过获取矩阵中的主成分,展开主成分分析 $B=\mathrm{PCA}(\boldsymbol{X})$,计算矩阵中每列的平均值 $\boldsymbol{M}=\mathrm{mean}(\boldsymbol{X})$,对原始数据矩阵 $\boldsymbol{X}$ 执行居中处理,即减去平均值,以获得一个新的矩阵 $\boldsymbol{X}'$,以消除不同传感指标之间的量纲差异和均值偏移:

$$\boldsymbol{X}'=\boldsymbol{X}-\boldsymbol{E}\boldsymbol{M}=\begin{bmatrix} \sigma_{11}^2 & \sigma_{11}^2 & \cdots & \sigma_{1m}^2 \\ \sigma_{11}^2 & \sigma_{22}^2 & \ddots & \vdots \\ \sigma_{m1}^2 & \sigma_{m2}^2 & \cdots & \sigma_{mm}^2 \end{bmatrix} \tag{8}$$

式中:$\sigma$ 为标准差;$\boldsymbol{E}$ 为 $n\times m$ 的全 1 矩阵,用于将平均值向量 $\boldsymbol{M}$ 扩展到与 $\boldsymbol{X}$ 相同的维度以便展开逐元素相减。$\boldsymbol{X}'$中每个元素为 $X_{ij}=X_{ij}-M_{ij}$($M_{ij}$为第 $j$ 列的第 $i$ 个平均值)。

根据居中处理后的矩阵 $\boldsymbol{X}'$,计算其协方差矩阵 $\boldsymbol{C}$[11],其中 $X_{ij}$间的协方差为:

$$\boldsymbol{C}=\frac{1}{n-1}\boldsymbol{X}'^{\mathrm{T}}\boldsymbol{X}' \tag{9}$$

式中:$\boldsymbol{X}'^{\mathrm{T}}$ 为 $\boldsymbol{X}'$的转置矩阵。协方差矩阵 $\boldsymbol{C}$ 的元素 $C_{ij}$ 为第 $i$ 个传感指标和第 $j$ 个传感指标之间的协方差,用于衡量它们之间的线性相关程度。协方差公式为:

$$C_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(X'_{ki}-\bar{X}_i)(X'_{kj}-\bar{X}_j) \tag{10}$$

式中:$\bar{X}_i$、$\bar{X}_j$ 为第 $i$ 列和第 $j$ 列的平均值(在居中处理后均为 0)。

通过对协方差矩阵 $\boldsymbol{C}$ 的特征值分解 $\boldsymbol{C}=\boldsymbol{W}\wedge\boldsymbol{W}\boldsymbol{T}$,获取特征值 $\lambda_i$ 及其对应的特征 $\omega_i(i=1,2,\cdots,m)$。这些特征值代表了各主成分所解释的方差量,特征向量则构成了特征向量矩阵 $\boldsymbol{W}$,同时形成一个对角矩阵$\wedge$[12]。基于特征值的大小排序,选取前 $k$ 个最大的特征值所对应的特征向量,即前 $k$ 个主成分。利用这些选定的特征向量,构造出一个新的矩阵 $\boldsymbol{W}_k=[\boldsymbol{\omega}_1,\boldsymbol{\omega}_2,\cdots,\boldsymbol{\omega}_k]$。

将经过居中处理的矩阵 $\boldsymbol{X}'$投影到特征向量矩阵 $\boldsymbol{W}_k$ 上,即可获得降维后的数据矩阵 $\boldsymbol{Y}=\boldsymbol{X}'\boldsymbol{W}_k$。该矩阵是一个 $n\times k$ 维度的矩阵,其列向量表示原始数据在各个主成分方向上的投影结果。

使用坡度图(或称为特征值图)来确定最佳的主成分个数 $k$。在坡度图中,选择曲线开始平缓之前的最后一个特征值对应的成分数作为 $k$。调整 $k$ 以获取最佳性能,提高降维处理的效果。由此,降维后的数据更加紧凑,特征之间的冗余信息被有效去除,使得分类器能够更加专注于区分不同类别的关键特征,从而提高了标注的准确性。

### 2.2 基于自适应 SVM 的传感数据动态语义标注

传感设备采集的数据元素间存在长距离依赖联系和短距离依赖联系,且这种依赖关系会随时间、环境等因素动态变化。针对这种动态特性,静态标注方法通常基于固定的特征关联规则进行标注,单一标注发生偏移的现象,数据的动态性使得数据特征不断更

新,这些规则不再适用,难以实时跟踪这些变化。

SVM 作为一种强大的分类器[13],在本文中被用于对降维后的传感数据进行分类。其基于最大间隔原理的分类策略使得 SVM 在处理非线性数据和复杂分类边界时具有显著优势。在动态数据环境下,获取大量标注数据往往成本高昂且耗时较长。通过结合未标注数据,可以充分利用数据中的有用信息,提高分类器的泛化能力。主动学习通过选择最有价值的未标注数据进行标注,提高了标注的效率和质量;半监督学习则利用标注数据和未标注数据之间的关系进行模型训练。在动态数据标注过程中,通过主动学习和半监督学习的迭代优化,分类器能够不断适应数据的动态变化。自适应 SVM 能够根据数据的动态性变化自适应调整分类器性能。在每一轮标注后,根据风险评估结果选择最优标注样本,进一步提高了标注的准确性和效率。完整流程如下:

①初始化

将 PCA 降维后的传感数据作为输入,划分为标注集 $\boldsymbol{L}\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_l,y_l)\}$(包含已标注的 $l$ 个传感数据及其特征向量 $\boldsymbol{x}$ 和标签 $y$),未标注集 $\boldsymbol{U}\{\boldsymbol{x}_{l+1},\boldsymbol{x}_{l+2},\cdots,\boldsymbol{x}_{l+u}\}$(包含未标注的传感数据及其特征向量 $\boldsymbol{x}$),输出为最终自适应的分类器 $f(x)$ 及标注结果。

②构建图与加权全矩阵

根据标注集 $\boldsymbol{L}$ 和未标注集 $\boldsymbol{U}$ 构建图 $H=(V,E)$,其中 $\boldsymbol{V}=\boldsymbol{L}\cup\boldsymbol{U}$,根据数据之间的相似性构建 $n\times n$ 的加权矩阵 $\boldsymbol{w}$,权值定义为:

$$\boldsymbol{w}_{ij}=\exp\left[-\sum_{d=1}^{n}\frac{(x_{id}-x_{jd})^2}{\sigma_d^2}\right] \tag{11}$$

式中:$n=l+u$;$l$ 为标注数据的数量;$u$ 为未标注数据的数量;$x_{id}$ 为数据点 $x_i$ 的第 $d$ 个元素;$\sigma_d^2$ 为第 $d$ 维的长度尺度参数。公式衡量了数据点 $x_i$ 和 $x_j$ 之间的相似性,相似性越高,$w_{ij}$ 的值越大。

③执行半监督学习标注

标注过程分为首轮、最后一轮和其他每轮三种情况。

A. 首轮和最后一轮标注:

从未标注集 $U$ 中选择 $u$ 个拥有最大 $d_{\text{A-SVM}}$(传感数据实例到 SVM 分类器的距离)的未标注样本,选取那些最可能包含有用信息的样本实施标注,公式为:

$$d_{\text{A-SVM}}(x_i)=\frac{1}{1+\exp[A-d(x_i,x_{\text{SVM}})]} \tag{12}$$

式中:$A$ 为根据训练数据估计的常数。

执行由 SVM 增强的半监督学习[14],从而使得标注过程更加灵活和智能,能够更好地应对复杂多变的传感数据标注需求。利用式(13)计算增强图的谐波函数,通过未标注数据来优化分类器,公式为:

$$h_u=[I-(1-\gamma)P_{uu}]^{-1}(1-\gamma)P_{ul}h_l+\gamma f_u \tag{13}$$

式中:$\gamma$、$1-\gamma$ 为 $x_i$ 到其适配结点和其他点的迁移概率;根据输出的 $h_u$,选择 $k$ 个具有最大 $h_u$ 的未标注传感数据作为下一轮标注的候选。

B. 其他每轮标注:

从未标注集 $\boldsymbol{U}$ 中选择 $u$ 个有最小 $|d_{\text{A-SVM}}|$(与 SVM 分类器距离最接近的样本)的未标注样本,以期望获得最大的信息增益。使用这些样本执行标准的半监督学习,利用下式计算未标注数据的标签:

$$h_u=(F_{uu}-E_{uu})^{-1}E_{ul}h_l=(I-P_{uu})^{-1}P_{ul}h_l \tag{14}$$

式中:$h_l$ 是函数 $g$ 在标注数据集 $\boldsymbol{L}$ 上的取值。

计算增入新标记实例 $(x_k,y_k)$ 后分类器 $f(x)$ 的风险变化 $R(f^{+x_k})$、$R(h^{+x_k})$,评估不同标注选择对分类器性能的影响:

$$X_k^*=\arg\min_{x_k\in F_u^i}\{\lambda R(f^{+x_k})+(1-\lambda)R(h^{+x_k})\} \tag{15}$$

根据风险评估的结果,选择 $k$ 个具有最小风险的未标注传感数据实施标注。

④迭代与终止

将上一步选出的 $k$ 个传感数据从未标注集 $\boldsymbol{U}$ 中移除,并加入到标注集 $\boldsymbol{L}$ 中。如果还有未标注数据,则重复步骤 3;否则,结束算法。

⑤输出最终分类器与标注结果

利用最终的标注集合 $\boldsymbol{L}$ 训练得出一个自适应分类器 $f(x)$,并输出此分类器及所有数据对应的标注结果。在每一轮的标注流程结束后,对标注集 $\boldsymbol{L}$ 和未标注集 $\boldsymbol{U}$ 实施更新,并将新标注的数据整合到训练集中,用以重新训练或更新分类器 $f(x)$。最终,输出的是一个经过优化的自适应分类器 $f'(x)$ 及其对应的标注结果。

## 3　实验与分析

实验选取某电力公司无线传感网络近一年的传感数据作为样本对象,每条记录均包含不同的关键属性,如光照强度、振动频率、电流、电压、温度、湿度、转速、位移、噪声水平以及压力等,并对应主要标签,包括正常状态、故障预警、紧急状态、稳定运行以及异常波动等,共包括 157 429 条数据。从总数据集中随机抽取 70% 的数据作为训练集,从剩余的 30% 数据中再随机抽取 15% 作为验证集,剩下的 15% 的剩余数据作为测试集。传感网络架构如图 2 所示。

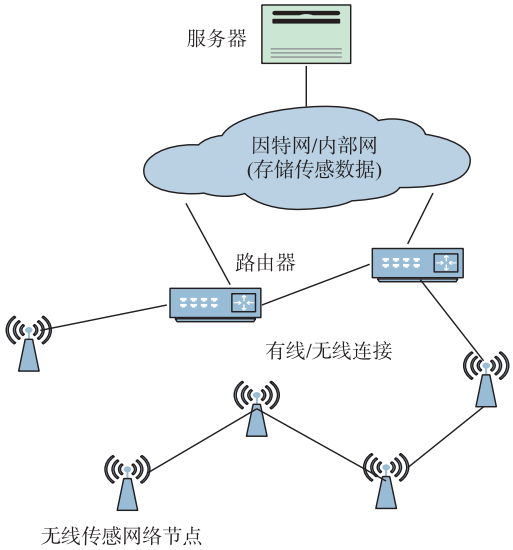


**图 2 传感网络环境**

在实验开始前,设定初始的实验参数:①数据平滑的窗口大小为 5;②线性插值法步长为 0.1;③数据归一化范围为[0,1];④PCA 降维主成分>0.1;⑤ASVM-SAL 方法标注集初始大小为 200;⑥每轮迭代标注样本数为 10;⑦最大迭代次数为 1 000 次。

为了进一步确保实验结果的稳定性和可靠性,采用 $K$ 折交叉验证方法对设置好参数的模型进行评估。选择 $K=10$,即将训练集和验证集合并后的数据平均分成 10 份,每份包含约 11 728 条数据。依次将其中 9 份作为新的训练集,剩下的 1 份作为验证集,进行模型训练和验证。例如,第一次将第 1~9 份作为训练集,第 10 份作为验证集;第二次将第 2~10 份作为训练集,第 1 份作为验证集,以此类推,共进行 10 次训练和验证过程。在每次训练和验证过程中,记录模型在验证集上的标注准确率,取平均值,最终结果为 97.69%,证明所提方法具有良好的应用性,避免了模型过拟合和欠拟合。

基于上述设置,选取测试集的原始传感数据,利用所设计的 PCA 方法对其实施降维处理。对比降维前后的数据维度变化,绘制维度变化对比,如图 3 所示。

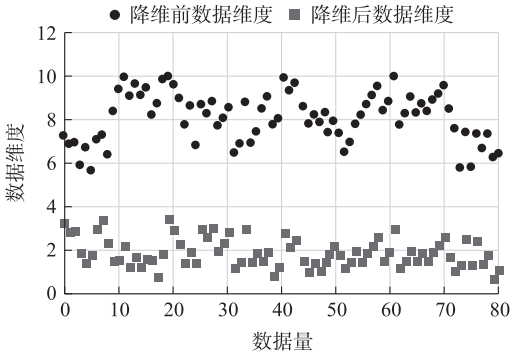


**图 3 所提方法数据降维结果**

实验结果显示,降维前的原始传感数据维度高达数十维,而通过 PCA 降维处理后,数据维度显著降低,而且更加集中,便于后续的分析和处理,验证了该方法可以为后续传感数据语义标注提供更精简的数据。

引入两种传统方法(动态句法剪枝机制[3]和 GoT 模型[5])作为对比方法,验证所提方法和这两种方法对传感数据动态语义标注的有效性,根据实验统计数据正确的真实标注类型(即真实标签)并与三种方法对比,验证三种方法的标注精度。

经过对比验证,如表 1 所示,相较于动态句法剪枝机制和 GoT 模型,所提方法在所有标注类型(真实标签)上的标注结果均正确无误,表现出色;而动态句法剪枝机制在“正常状态”和“紧急状态”上出现了错误标注,GoT 模型则在“故障预警”和“稳定运行”上出现了误判;这两种传统方法在复杂多变的传感数据面前难以保证标注的精度和可靠性。

**表 1 各方法语义标注结果**

| 标注类型(真实标签) | 所提方法 | 动态句法剪枝机制 | GoT 模型 |
|---|---|---|---|
| 正常状态 | 正常状态 | 故障预警(错误) | 正常状态 |
| 故障预警 | 故障预警 | 故障预警 | 紧急状态(错误) |
| 紧急状态 | 紧急状态 | 异常波动(错误) | 紧急状态 |
| 稳定运行 | 稳定运行 | 稳定运行 | 正常状态(错误) |
| 异常波动 | 异常波动 | 稳定运行(错误) | 异常波动 |

从量化分析角度对比三种方法的有效性,引入宏平均值 $F_{1\text{Macro}}$ 指标,评估三种方法的标注效果,计算公式为:

$$F_{1\text{Macro}}=\sum_{K\in O}F_1(K)/|\boldsymbol{O}| \tag{16}$$

式中:$F_1(K)$ 为标签 $K$ 对应的 $F_1$ 值;$\boldsymbol{O}$ 为整体标签集合。$F_{1\text{Macro}}$ 值越高表明语义标注方法的有效性和准确性越高,在样本百分比不断增加的条件下,得到三种方法的 $F_{1\text{Macro}}$ 值,如图 4 所示。

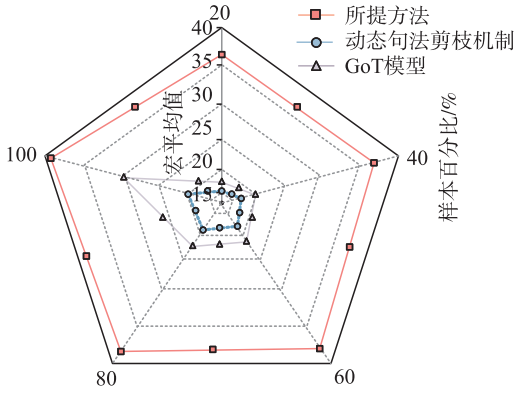


**图 4 各方法 $F_{1\text{Macro}}$ 值评估结果**

从图 4 的评估结果可以看出,所提方法在所有样本百分比条件下均表现出最高的宏平均值,平均为 36.8,充分证明了其标注的有效性和准确性;动态句法剪枝机制和 GoT 模型的宏平均值较低,且随着样本百分比的增加,其标注效果出现了不同程度的降低。表明这两种传统方法在复杂传感数据面前存在局限性,难以应对大规模、多样化的数据标注需求。这是因为所提方法考虑了传感数据的动态性,通过自适应 SVM 的半监督主动学习方法,实现了对传感数据的实时跟踪和动态标注。这种适应性使得方法能够在数据发生变化时及时调整分类策略,保持标注的高准确性。

## 4　结论

文章提出融合 PCA 与 SVM 的传感数据动态语义标注方法。首先,数据预处理步骤有效清洗了原始数据中的噪声和异常值,为后续处理奠定了坚实基础。接着,利用 PCA 降维技术,降低了数据的维度,不仅大幅减少了计算复杂度,还最大限度地保留了数据的关键信息。在此基础上,利用自适应 SVM 结合半监督主动学习,根据数据动态性自适应调整,避免单一标注偏移,使其能够更准确地捕捉传感数据的动态特征,实现了对数据的精准标注。这三个方法的综合运用,不仅显著提升了传感数据处理的效率与准确性,更为数据的深度挖掘和高级应用提供了强有力的支持。

**参考文献:**


[1] 张雪英,张春菊,汪陈,等. 面向中文文本的地质语义信息标注与语料库构建[J]. 高校地质学报,2023,29(3):429-438.

[2] 胡青宁,董金平,李婷玉,等. 基于关键字的数据元语义描述方法[J]. 东北石油大学学报,2022,46(3):107-114,12.

[3] 费豪,姬东鸿,任亚峰.基 于动态句法剪枝机制的中文语义角色标注[J]. 计算机学报,2022,45(8):1746-1764.

[4] Xu H,Sun H. Application of Rough Concept Lattice Model in Construction of Ontology and Semantic Annotation in Semantic Web of Things[J]. Scientific Programming,2022,2022(1):7207372.

[5] Xu Y,Zhao X,Jiang Z,et al. Intelligent Semantic Annotation for Mobile Services for IoT Computing from Heterogeneous Data[J]. Mobile Networks and Applications,2023,28(1):348-358.

[6] Liu J Z,Liu Y Z. High-Throughput Real-Time Reliable Data Collection in Wireless Sensor Network: Implementation and Analysis[J]. IEEE Sensors Journal,2024,24(17):28251-28266.

[7] 靳宗帅,张恒旭. 电力系统强噪声测量数据快速鲁棒回归平滑降噪[J]. 电力系统自动化,2023,47(3):169-176.

[8] 姜云木,阮鑫鑫,刘章军,等. 基于工程特性的主余震型地震动降维模拟[J]. 地震工程学报,2023,45(3):735-741.

[9] 杨启帆,段大卫,李楠,等. 基于主成分分析的串联电池组故障诊断实用方法[J]. 电力自动化设备,2022,42(12):210-216.

[10] 葛焌迟,赵为华. 矩阵数据基于鲁棒主成分分析的距离加权判别分析[J]. 计算机应用,2024,44(7):2073-2079.

[11] 栗铭婉,王超,李欣蔚,等. 基于主成分分析法的牵引供电系统负序电压责任量化方法[J]. 中国测试,2023,49(S2):65-69.

[12] 黎力韬,王龙林,王华. 基于 IMA-SVM 的斜拉桥索力可靠度计算模型[J]. 计算机仿真,2024,41(7):319-325.

[13] Guo J Q, Wu H Y, Chen X L, et al. Adaptive SV-Borderline SMOTE-SVM Algorithm for Imbalanced Data Classification[J]. Applied Soft Computing, 2024, 150: 110986.

[14] 高泽鋆,曹菲,何川,等. 基于半监督学习网络的雷达有源干扰识别[J]. 探测与控制学报,2022,44(6):93-101.


**张书月**(1985—),女,汉族,湖北襄阳人,硕士,副教授,主要研究方向:数据挖掘、计算机应用,zshuyue1231@163.com;

**段　超**(1987—),男,汉族,甘肃庆阳人,硕士,副教授,主要研究方向:移动通信,通感一体化和人工智能,duanchao@gzgs.edu.cn。