DOI: 10.11992/tis.201804051
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180611.1014.002.html

多标记学习自编码网络无监督维数约简

杨文元
（闽南师范大学 福建省粒计算及其应用重点实验室，福建 漳州 363000）

摘　要：多标记学习是针对一个实例同时与一组标签相关联而提出的一种机器学习框架，是该领域研究热点之一，降维是多标记学习一个重要且具有挑战性的工作。针对有监督的多标记维数约简方法，提出一种无监督自编码网络的多标记降维方法。首先，通过构建自编码神经网络，对输入数据进行编码和解码输出；然后，引入稀疏约束计算总体成本，使用梯度下降法进行迭代求解；最后，通过深度学习训练获得自编码网络学习模型，提取数据特征实现维数约简。实验中使用多标记算法 ML-kNN 做分类器，在 6 个公开数据集上与其他 4 种方法对比。实验结果表明，该方法能够在不使用标记的情况下有效提取特征，降低多标记数据维度，稳定提高多标记学习性能。

关键词：多标记学习；维数约简；无监督学习；神经网络；自编码器；机器学习；深度学习；特征提取
中图分类号：TP183　　**文献标志码**：A　　**文章编号**：1673−4785(2018)05−0808−10

中文引用格式：杨文元. 多标记学习自编码网络无监督维数约简[J]. 智能系统学报, 2018, 13(5): 808–817.
英文引用格式: YANG Wenyuan. Unsupervised dimensionality reduction of multi-label learning via autoencoder networks[J]. CAAI transactions on intelligent systems, 2018, 13(5): 808–817.

Unsupervised dimensionality reduction of multi-label learning via autoencoder networks

YANG Wenyuan
(Fujian Key Laboratory of Granular Computing and Application, Minnan Normal University, Zhangzhou 363000, China)

Abstract: Multi-label learning is a machine learning framework that simultaneously deals with data associated with a group of labels. It is one of the hot spots in the field of machine learning. However, dimensionality reduction is a significant and challenging task in multi-label learning. In this paper, we propose a unsupervised dimensionality reduction method for supervised multi-label dimensiona reduction methods via autoencoder networks. Firstly, we build the autoencoder neural network to encode the input data and then decode them for output. Then we introduce the sparse constraint to calculate the overall cost, and further use the gradient descent method to iterate them. Finally, we obtain the autoencoder network learning model by deep learning training, and then extract data features to reduce dimensionality. In the experiments, we use a multi-label algorithm (ML-kNN) as the classifier, and compare them with four other methods on six publicly available datasets. Experimental results show that the proposed method can effectively extract features without using label learning; thus, it reduces multi-label data dimensionality and steadily improves the performance of multi-label learning.

Keywords: multi-label learning; dimensionality reduction; unsupervised learning; neural networks; autoencoder; machine learning; deep learning; feature extraction

收稿日期：2018−04−25.　**网络出版日期**：2018−06−11.
基金项目：国家自然科学青年基金项目 (61703196)；福建省自然科学基金项目 (2018J01549).
通信作者：杨文元. Email: yangwy@mnnu.edu.cn.

真实世界中的对象通常具有不止一种语义标记，经常表现为多义性，即一个对象可能与多个类别标记相关联。如一幅海边景色图片可以同时

标注“蓝天”“白云”“大海”“沙滩”等语义标记, 对于多个标记对象的处理方式是为每个图像赋予一个标记子集, 并进行建模和学习, 这就形成多标记学习框架[1]。在多标记学习框架下, 每个示例由对应的多个标记构成的特征向量进行描述, 学习的目标是将多个适当的标记赋予需要预测的未知示例[2]。

随着信息化的快速发展, 数据和资源呈海量特征, 数据的标注结构复杂程度也不断增加, 单标记方法无法满足分析处理要求[1], 以机器学习技术为基础的多标记学习技术现已成为一个研究热点, 其研究成果广泛地应用于各种不同领域, 如图像视频的语义标注、功能基因组、音乐情感分类以及营销指导等[3]。

在多标记学习过程中, 高维数据训练和预测都需要更多的计算时间和空间。降维减少了特征数却提高了算法效率和学习性能, 可避免过拟合现象和过滤掉冗余特征[4-6]。因此, 降低数据的维度具有重要意义。

高维数据降维, 主要有线性维数约简和非线性维数约简两种方法。线性维数约简的方法有主成分分析方法 (principal component analysis, PCA)、独立成分分析方法 (independent component correlation algorithm, ICA)、线性判别分析法 (linear discriminant analysis, LDA) 和局部特征分析法 (local feature analysis, LFA) 等。非线性降维方法有等距特征映射方法 (isometric feature mapping, ISOMAP) 和局部线性嵌入方法 (locally linear embedding, LLE) 等[7-8]。多标记学习的有监督降维方法有依赖最大化 (MDDM) 算法[5], 半监督方法主要是采用联合降维方法[9]和依赖最大化方法[10]。

与一般多标记有监督的降维方法不同, 提出一种自编码网络的无监督多标记维数约简方法 (multi-label unsupervised dimensionality reduction via autoencoder networks, MUAE), 首先构建自编码神经网络, 仅使用特征数据作为输入, 进行编码和解码输出以提取特征, 在处理过程引入稀疏约束并将输出数据与输入数据对比, 计算总体成本误差, 应用梯度下降法进行迭代更新, 通过深度学习训练获得自编码网络学习模型, 提取数据特征, 最后以多标记学习 ML-kNN 算法作为统一的分类评价基准, 并在 6 个公开数据集上与其他 4 种方法对比。实验结果表明, 该方法能够在无监督情况下有效提取特征, 降低多标记数据维度, 得到较好的学习效果。

1　多标记学习

多标记的样本由一个示例和对应的多个标记构成, 多标记学习是一种机器学习框架[1-2]。下面的内容, 简要地介绍多标记学习的问题定义和学习算法。

1.1　多标记问题定义

假设$\boldsymbol{X} \in \mathbf{R}^d$代表$d$维的示例空间, $\boldsymbol{Y}=\{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_l\}$代表包含$l$个类别的标记空间。给定多标记训练集$\boldsymbol{D}=\{(\boldsymbol{x}_i, \boldsymbol{Y}_i) \mid 1 \leqslant i \leqslant m\}$, 其中$\boldsymbol{x}_i$为$d$维的属性向量, 即$\boldsymbol{x}_i=[x_{i1}\ x_{i2} \cdots x_{id}]^{\mathrm{T}}$, 而$\boldsymbol{Y}_i \in \boldsymbol{Y}$为与$\boldsymbol{x}_i$对应的一组类别标记, 多标记学习的目标是从中学习得到一个多标记分类器$g: \boldsymbol{X} \to 2^{\boldsymbol{Y}}$。因此, 对于所有示例$\boldsymbol{x} \in \boldsymbol{X}$, 分类器预测隶属于该示例的类别标记集合为$g(\boldsymbol{x}) \subseteq \boldsymbol{Y}$, 称多标记学习[1]。

多标记的集合空间如果太大则会造成学习困难, 因此需要充分利用标记之间的“相关性”来辅助学习过程的进行。基于考察标记之间相关性的不同方式, 多标记学习问题求解策略有三类[1, 11]。

“一阶 (first-order)”策略: 只考察每一个单个标记, 不考虑标记之间的相关性, 将多标记学习问题分解为多个独立的二分类问题。该策略实现简单, 效率较高, 但学习的泛化性能不高[11]。

“二阶 (second-order)”策略: 考察两两标记之间的相关性和交互关系, 该类方法的泛化性能较好, 但不能很好处理多标记间的二阶以上相关性[11]。

“高阶 (high-order)”策略: 考察任一标记对其它所有标记的影响以及一组随机标记集合的相关性等。该类策略较好地反映了真实世界的标记相关性, 但复杂度一般过高, 难以处理大规模学习问题[11]。

1.2　多标记学习算法

目前已经涌现出了大量的多标记学习算法, 可以分为问题转换和算法适应两类方法[1]。

问题转换方法的基本思想是将多标记学习问题转换为其他已知的学习问题进行求解, 代表性学习算法有 Binary Relevance、Calibrated Label Ranking 和 Random k-labelsets。算法适应方法的基本思想是通过对常用监督学习算法进行改进, 将其直接用于多标记学习, 代表性学习算法有 ML-kNN, Rank-SVM 和 LEAD[1]。

ML-kNN 算法[12]是对 k 近邻 (k-nearest neighbors, kNN) 算法进行改造以适应多标记数据分类, 算法的基本思想是采用 kNN 分类准则, 统计近邻样本的类别标记信息, 通过最大化后验概率的方式推理未知示例的标记集合。

2 自编码网络

自编码网络包含数据输入层、隐藏层、输出重构层。如图 1 所示，自编码器由编码器 (encoder) 和解码器 (decoder) 两部分构成。其作用是将输入样本压缩到隐藏层，然后解压，在输出层重建样本[13-15]。

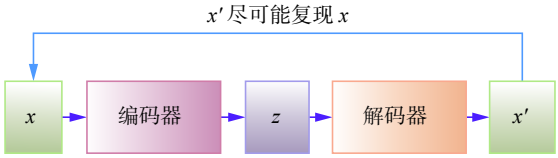


图 1 自编码网络模型

Fig. 1 Model of autoencoder networks

自编码网络是一种不需要标记的无监督学习模型，它试图学习一个函数$h(\boldsymbol{x})\approx\boldsymbol{x}$,训练网络使得输出逼近输入，也就是每个样本$\boldsymbol{x}$的学习目标也是$\boldsymbol{x}$，这样自编码器自己生成标签，而且标签就是样本数据本身，所以也称为自监督学习或自学习。如图 1 所示，从输入x通过编码器到$\boldsymbol{z}$，然后经解码器到$\boldsymbol{x}'$，自编码器的目的是，让输出$\boldsymbol{x}'$尽可能复现输入$\boldsymbol{x}$。系统的输出$\boldsymbol{x}'$能够复原原始数据$\boldsymbol{x}$，说明$\boldsymbol{z}$的维度虽然与$\boldsymbol{x}$的维度不同，但承载了原始数据的所有信息，只是形式不同，是已经变换特征的某种形式。如果对隐含层进行约束使得$\boldsymbol{z}$的维度小于$\boldsymbol{x}$的维度，就可以实现无监督数据降维[16]。

自编码器 AE(autoencoder) 接受一个输入$\boldsymbol{x}\in\mathbf{R}^m$，通过一个编码器将它映射到一个隐藏层，用$\boldsymbol{z}$表示，如式 (1) 所示：

$$\boldsymbol{z}=s(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b}) \tag{1}$$

通常称$\boldsymbol{z}$为编码，也称为潜在变量或潜在表示，s是一种基于元素的激活函数，可以是 sigmoid 函数或双曲正切函数，$\boldsymbol{W}$是一个权重矩阵，$\boldsymbol{b}$是一个偏差向量。

在自编码器的解码阶段，解码器将$\boldsymbol{z}$映射到与$\boldsymbol{x}$形状相同的重构$\boldsymbol{x}'$，即

$$\boldsymbol{x}'=s'(\boldsymbol{W}'\boldsymbol{z}+\boldsymbol{b}') \tag{2}$$

一般而言，解码器中的s'、$\boldsymbol{W}'$和$\boldsymbol{b}'$可能与编码器中的s、$\boldsymbol{W}$和$\boldsymbol{b}$不同，主要取决于自编码器的设计，自编码器的学习函数为$h(\boldsymbol{x})=s\circ s'$。

重建误差可以用许多方法测量，可根据给定输入的分布假设而制定。一般可以采用样本的代价函数，单个样本的代价函数为

$$c(\boldsymbol{W},\boldsymbol{b},\boldsymbol{x}_i,\boldsymbol{x}_i')=\frac{1}{2}\left\|\boldsymbol{x}_i'-\boldsymbol{x}_i\right\|^2 \tag{3}$$

m个样本的整体平方误差成本函数为

$$C(\boldsymbol{W},\boldsymbol{b})=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left\|\boldsymbol{x}_i'-\boldsymbol{x}_i\right\|^2+\frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s+1}(\boldsymbol{W}_{ji}^{l}) \tag{4}$$

式中：m是样本个数；n_l是网络层数；n_s是l层的神经元数量；λ是权重衰减参数，用于控制公式中两项的相对重要性。$C(\boldsymbol{W},\boldsymbol{b})$是整体样本代价函数，第一项是均方差项，第二项是权重衰减项或称为规则化项，其目的是防止过度拟合。

训练目标就是获得总体样本成本误差$C(\boldsymbol{W},\boldsymbol{b})$最小的$\boldsymbol{W},\boldsymbol{b}$，即

$$\arg\min_{\boldsymbol{W},\boldsymbol{b}}(C(\boldsymbol{W},\boldsymbol{b})) \tag{5}$$

如果输入是完全随机的，每个输入数据都独立于其他特征的高斯分布，则编码器的压缩任务将非常困难。但是，如果数据中存在结构，有些输入要素是相关或有冗余，则该算法将能够发现一些相关性。自动编码器通常最终会学习与 PCA 非常相似的低维表示。事实上，如果每两层之间的变换均为线性且训练误差是二次型误差时，该网络等价于 PCA。而自编码网络使用非线降维，更符合数据的实际情况，这种机制使得其效果比 PCA 更优。

自编码网络可以实现无监督的自我学习，把这种自我学习扩展到深度学习网络，即拥有多个隐藏层的神经网络，以提取多标记的数据特征，实现多标记学习的无监督维数约简。

3 多标记维数约简的自编码网络方法

3.1 多标记维数约简

多标记学习与单标记学习一样面临“维度灾难”的挑战，所以维数约简结果的好坏直接影响着分类器的精度和泛化性能，特别是对于基因序列、图像处理等高维数据，影响更加显著。数据维度过大，不仅会增加计算时间和空间的复杂度，还会降低多标记学习性能。如果在多标记学习训练之前，通过一定的特征选择或提取方法，去掉不相关或冗余属性，反而可以获得更令人满意的知识学习模型[17-18]。降低高维数据的维度是多标记学习中一个重要的研究课题，很多学者研究多标记数据的降维方法以提高多标记学习算法的效果[19]。

已有的多标记数据维数约简方法可以分为两大类：特征选择 (feature selection) 和特征提取 (feature extraction)。特征选择是给定一个多标记分类算法和训练集，通过优化某个多标记损失函数对属性子集进行评价，选择使损失达到最小的属性子集作为最终结果[20]。而特征提取是通过空间变

换, 将某些原始特征映射到其他低维空间, 生成一些新的特性[3, 5], 特征提取后的新特征是原来特征的一个变换映射, 不是原特征一个子集。

3.2　多标记学习维数约简的自编码网络方法

基本自编码网络可以解决数量很小的隐藏单元问题, 而高维数据的隐藏单元数量很大, 为此, 对隐藏单元进行稀疏约束, 使得自编码器可以从大量的隐藏单元中发现高维数据中的相关结构, 提取关键特征, 实现维数约简。自编码网络, 由输入层、隐含层1,2,⋯,i,⋯n和输出层组成, 如图 2 所示。

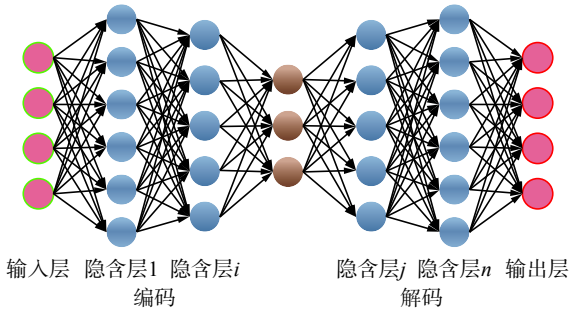


图 2　自编码网络

Fig. 2　Autoencoder network

如果激活函数采用 sigmoid 函数, 当神经元的输出值接近于 1, 则神经元是“活动的”, 如果它的输出值接近于 0, 则神经元是“无效的”。

令a_j表示自编码器中激活隐藏单元j, 当网络被赋予特定的输入$\boldsymbol{x}$时, 用$a_j(\boldsymbol{x})$表示$\boldsymbol{x}$激活隐藏单元j的输出值。如果输入$\boldsymbol{x}$有m个样本, 则隐藏单元j的平均激活为$\bar{\rho}_j$, 即

$$\bar{\rho}_j = \frac{1}{m}\sum_{i=1}^{m} a_j(\boldsymbol{x}_i) \tag{6}$$

用稀疏参数ρ对每个隐藏神经元的平均激活进行稀疏约束, ρ的取值接近于零, 比如取 0.05。为满足这个约束, 引入伯努利随机变量 KL(Kull-back-Leibler) 散度来度量平均值$\bar{\rho}_j$和ρ之间的距离, 即

$$\mathrm{KL}(\rho\|\bar{\rho}_j) = \rho\log\frac{\rho}{\bar{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\bar{\rho}_j} \tag{7}$$

隐藏层中所有神经元的 KL 散度之和作为优化目标的处罚项, 以惩罚显著偏离ρ, 即

$$\sum_{j=1}^{n_s}\mathrm{KL}(\rho\|\bar{\rho}_j) \tag{8}$$

式中: n_s是隐藏层中所有神经元的数量。如果$\bar{\rho}_j=\rho$, 那么$\mathrm{KL}(\rho\|\bar{\rho}_j)=0$; 否则$\bar{\rho}_j$会随着偏离$\rho$而单调增加, 所以只要使这个惩罚项最小化, 就会导致$\bar{\rho}_j$接近于ρ。

稀疏约束后, 样本的总体成本为

$$C_{\mathrm{sparse}}(\boldsymbol{W},\boldsymbol{b}) = C(\boldsymbol{W},\boldsymbol{b}) + \beta\sum_{j=1}^{n_s}\mathrm{KL}(\rho\|\bar{\rho}_j) \tag{9}$$

式中β是控制稀疏惩罚项的权重。

稀疏约束后训练目标也是总体成本误差最小, 即

$$\arg\min_{\boldsymbol{W},\boldsymbol{b}}(C_{\mathrm{sparse}}(\boldsymbol{W},\boldsymbol{b})) \tag{10}$$

为了求解上述总体成本误差最小问题, 采用反向传播算法计算成本偏导数, 先求出成本函数$C_{\mathrm{sparse}}(\boldsymbol{W},\boldsymbol{b})$对$\partial\boldsymbol{W}_{ij}^{l}$的偏导数, 得到

$$\frac{\partial}{\partial W_{ij}^{l}}C_{\mathrm{sparse}}(\boldsymbol{W},\boldsymbol{b}) = \frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial W_{ij}^{l}}C(\boldsymbol{W},\boldsymbol{b},\boldsymbol{x}_i,\boldsymbol{x}_i') + \lambda W_{ij}^{l} \tag{11}$$

以及对$\boldsymbol{b}_i^l$求偏导数, 得到

$$\frac{\partial}{\partial \boldsymbol{b}_i^{l}}C_{\mathrm{sparse}}(\boldsymbol{W},\boldsymbol{b}) = \frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial \boldsymbol{b}_i^{l}}C(\boldsymbol{W},\boldsymbol{b},\boldsymbol{x}_i,\boldsymbol{x}_i') \tag{12}$$

采用梯度下降迭代法, 按公式 (13)、(14) 进行迭代:

$$\boldsymbol{W}_{ij}^{l} = \boldsymbol{W}_{ij}^{l} - \alpha\frac{\partial}{\partial \boldsymbol{W}_{ij}^{l}}C_{\mathrm{sparse}}(\boldsymbol{W},\boldsymbol{b}) \tag{13}$$

$$\boldsymbol{b}_i^{l} = \boldsymbol{b}_i^{l} - \alpha\frac{\partial}{\partial \boldsymbol{b}_i^{l}}C_{\mathrm{sparse}}(\boldsymbol{W},\boldsymbol{b}) \tag{14}$$

综合上述推导过程和结果, 设计多标记学习的自编码网络无监督降维算法 MUAE 如下。

算法　MUAE

输入　$\boldsymbol{X}^{n\times m}$特征矩阵, $\boldsymbol{Y}^{n\times p}$标记矩阵, 约简维度$d$。

输出　多标记学习分类结果。

1) 初始化权重矩阵$\boldsymbol{W}^l$和偏离向量$\boldsymbol{b}^l$;

2) for epoch =1:k;

3) 计算样本总成本$C_{\mathrm{sparse}}(\boldsymbol{W},\boldsymbol{b})$; //由式 (9) 计算;

4) 迭代权重矩阵$\boldsymbol{W}^l$和偏离向量$\boldsymbol{b}^l$; //由式 (13) 和式 (14) 计算;

5) end for

6) 返回输出层特征矩阵$\boldsymbol{X}^{n\times d}$;

7) 多标记算法 ML-kNN($\boldsymbol{X}^{n\times d}$, $\boldsymbol{Y}^{n\times p}$) 计算分类结果//按文献[12]计算。

4　实验结果与分析

4.1　实验数据与对比算法

多标记学习数据降维实验采用公开数据集[21], 各数据集的训练和测试样本数、标记数量与数据特征数量等基本情况如表 1 所示, 表中 6 个数据集 Arts、Business、Computers、Health、Recreation、Reference 的名称前面分别用 A、B、C、D、E、F 对应标注, 以方便后续表格使用。

表 1　数据集基本描述

Table 1　Data description

数据集/数据信息	训练样本数	测试样本数	标记个数	特征个数
A: Arts	2 000	3 000	26	462
B: Business	2 000	3 000	30	438
C: Computers	2 000	3 000	33	681
D: Health	2 000	3 000	32	612
E: Recreation	2 000	3 000	22	606
F: Reference	2 000	3 000	33	793

实验过程中，将 MUAE 算法与 4 种算法进行对比，对比算法分别是线性维数约简主成分分析 PCA 算法[22]、非线性维数约简局部保留投影 LPP[23] 算法和拉普拉斯特征映射 LE 算法[24]，以及多标记依赖最大化 MDDM 算法[5]。

在维数约简后统一使用 ML-kNN 算法[12]进行多标记分类，其中$k=10$，并以 ML-kNN 算法在原始特征空间的评价性能作为参照基线，记为 Baseline。MDDM 算法的$\mu=0.5$，LLP 算法分类时构造邻接图的最近邻个数与 ML-KNN 算法一样设置$k=10$。所有维数约简方法降维到相同的维度进行对比，所有算法在 6 个数据集上的特征降维百分比为 10%、20%、30%、40%、50%、60%、70%、80%、90%、100%，共 10 个百分比的实验对比。

4.2　多标记学习评价指标

在多标记学习问题中，由于每个对象可能同时具有多个类别标记，因此传统监督学习中常用的单标记评价指标无法直接用于多标记学习系统的性能评价。因此，研究者们相继提出了一系列多标记评价指标，一般可分为两种类型，即基于样本的评价指标 (example-based metrics)[25]以及基于类别的评价指标 (label-based metrics)[26]。本文主要采用 5 种指标，即平均精度 (average precision)、汉明损失 (Hamming loss)、排名损失 (ranking loss)、一错误 (oneerror) 和覆盖 (coverage)，具体的计算公式如下。

1) 平均精度

$$\text{Averageprecision}(h)=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i|}\sum_{y\in Y_i}\frac{\left|\{y'|\text{rank}_f(x_i,y')\leqslant\text{rank}_f(x_i,y),y'\in Y_i\}\right|}{\text{rank}_f(x_i,y)} \tag{15}$$

平均精度，是一种最直观的评价方式，评价样本的类别标记排序序列中，排在相关标记之前的标记占标记集合的平均比例，这个指标是相关标记预测的概率平均。

2) 汉明损失

$$\text{Hammingloss}(h)=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{q}|h(x_i)\Delta Y_i| \tag{16}$$

汉明损失，是通过计算多标记分类器预测的标记结果与实际的标记差距来度量多标记分类器的性能。

3) 排名损失

$$\text{Rankingloss}(f)=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i||\bar{Y}_i|}\left|\{(y_1,y_2)|f(x_i,y_1)\leqslant f(x_i,y_2),(y_1,y_2)\in Y_i\times\bar{Y}_i\}\right| \tag{17}$$

排名损失，评价所有样本的预测标记排名中，不相关标记在相关标记前面的概率平均。

4) 一错误

$$\text{Oneerror}(f)=\frac{1}{p}\sum_{i=1}^{p}\arg\max_{y\in Y}f(x_i,y)\notin Y_i \tag{18}$$

一错误，该指标评价每个样本的预测标记排名中，排在第一位的标记不在该样本的相关标记集中的概率评价。

5) 覆盖

$$\text{Coverage}(f)=\frac{1}{p}\sum_{i=1}^{p}\max_{y\in Y_i}\text{rank}_f(x_i,y)-1 \tag{19}$$

覆盖，该指标评价每个样本的预测标记排名中需要在标记序列表中最少查找到第几位才可以找出所有与该样本相关的标记。

4.3　实验结果

5 种算法和基准算法共 6 个算法，在 6 个多标记数据集上用 5 个评价指标进行对比实验，实验的结果展示在表 2~6。

表 2　不同降维方法的平均精度

Table 2　Average precision of different algorithms

(↑)	A	B	C	D	E	F
Baseline	0.509 3	0.879 8	0.633 4	0.681 2	0.454 4	0.619 3
PCA	0.533 4	0.876 2	0.659 2	**0.723 3**	0.525 9	**0.663 9**
LPP	0.516 3	0.874 6	0.630 9	0.692 6	0.487 8	0.635 6
LE	0.496 3	0.865 6	0.621 2	0.686 9	0.456 0	0.613 1
MDDM	**0.557 5**	0.886 9	0.673 0	0.700 4	0.510 6	0.654 6
MUAE	0.535 9	**0.889 3**	**0.678 7**	0.715 3	**0.529 3**	0.662 6

表 2 是评价平均精度指标的实验结果，其数值是越高越好，最好的结果用黑体表示。实验结果显示，MUAE 方法在 Business、Computers、Recreation 三个数据集上都取得最好结果，PCA 在 Health

和 Reference 数据集上取得最好结果, MDDM 方法在 Arts 数据集上取得最好结果。能够在平均精度取得好的实验结果, 这是由于自编码深度网络能够通过自学习有效提取数据特征, 在有监督的多标记数据集上能够通过无监督方法取得好的降维效果。

另外, 4 个评价指标分别是汉明损失、排名损失、一错误和覆盖, 指标数值越小越好, 表 3~6 分别是这 4 种评价指标的各算法的实验结果, 最好的结果用黑体表示。MUAE 算法取得最好数据集个数分别为 3、3、2、2, 4 个表中的实验结果显示出 MUAE 方法总体上比其他 4 种算法和基准算法好。

表 3　不同降维方法的汉明损失

Table 3　Hamming loss of different algorithms

(↓)	A	B	C	D	E	F
Baseline	0.061 2	0.026 9	0.041 2	0.045 8	0.061 8	0.031 4
PCA	0.059 2	0.027 9	0.037 4	**0.039 2**	0.058 4	**0.027 6**
LPP	0.060 3	0.027 2	0.040 8	0.042 7	0.059 2	0.030 6
LE	0.061 7	0.028 2	0.042 1	0.043 3	0.061 0	0.031 6
MDDM	**0.058 5**	0.026 2	0.037 3	0.040 8	0.059 6	0.029 1
MUAE	0.059 3	**0.026 0**	**0.036 5**	0.040 2	**0.058 1**	0.028 3

表 4　不同降维方法的排名损失

Table 4　Ranking loss of different algorithms

(↓)	A	B	C	D	E	F
Baseline	0.152 0	0.037 4	0.092 2	0.060 5	0.191 2	0.091 9
PCA	0.141 4	0.041 5	0.085 9	**0.054 3**	0.165 7	**0.076 2**
LPP	0.148 7	0.043 2	0.092 7	0.060 7	0.187 5	0.087 0
LE	0.156 5	0.045 5	0.096 5	0.064 2	0.197 8	0.092 4
MDDM	**0.132 6**	**0.034 7**	0.082 6	0.055 7	0.169 8	0.078 1
MUAE	0.140 5	0.038 7	**0.080 6**	0.055 4	**0.164 9**	**0.076 2**

综合数据降维的各方法表现, 利用自编码进行无监督特征提取, 比无监督算法能够取得更好的效果, 这应该得益于自编码的思想和设计结构, 其能更好地表示输入数据的特征, 所以取得好的实验结果。

为了进一步分析自编码在不同降维百分比的性能, 以维度数量的 10% 开始, 步长以 10% 递增至 100%, 共 10 组, 结果以图的形式展示。图 3 是平均精度随特征降维百分比变化关系, MUAE 在 6 个数据集上比其他算法能取得更高的精度。图 3 还显示出平均精度在各个百分比的情况下, MUAE 算法精度高且很平稳, 没有大幅度变化, 而 LPP 和 LE 这两个算法随着降维百分比的增加精度反而逐步下降。

表 5　不同降维方法的一错误

Table 5　One-error of different algorithms

(↓)	A	B	C	D	E	F
Baseline	0.632 7	0.121 3	0.436 7	0.420 7	0.706 7	0.473 0
PCA	0.588 3	0.122 7	0.412 0	**0.355 0**	**0.606 7**	0.424 3
LPP	0.627 0	0.120 0	0.442 3	0.393 3	0.651 7	0.452 3
LE	0.649 8	0.135 0	0.452 0	0.380 7	0.695 0	0.475 7
MDDM	**0.555 7**	0.118 3	**0.392 7**	0.394 0	0.626 3	0.435 7
MUAE	0.591 0	**0.117 1**	0.412 0	0.364 0	0.619 7	**0.423 6**

表 6　不同降维方法的覆盖

Table 6　Coverage of different algorithms

(↓)	A	B	C	D	E	F
Baseline	5.445	2.185	4.416	3.305	5.097	3.542
PCA	5.189	2.315	4.220	**3.080**	4.507	**2.999**
LPP	5.368	2.377	4.492	3.265	5.036	3.361
LE	5.584	2.468	4.599	3.477	5.225	3.559
MDDM	**4.871**	2.071	**3.986**	3.096	4.578	3.071
MUAE	5.195	**2.065**	4.175	3.092	**4.501**	3.009

另外, 除数据集 Business 外的其他 5 个数据集, 所有算法在特征降维百分比较小的情况下, 平均精度的实验结果都比 Baseline 的结果好, 这表明大部分数据集确实存在冗余的特征, 各算法提取关键特征而去除了冗余特征, 因此, 多标记数据降维后, 学习精度反而得到不同程序的提高。

其余 4 个指标, 即汉明损失、排名损失、一错误和覆盖, 随特征降维百分比变化关系展示在图 4~7 中, 这 4 种指标越小越好。从图 4~7 显示的指标性能, 总体上 MUAE 方法比其他 4 种方法好, 曲线平稳, 起伏变化较小, 显示出 MUAE 方法稳定性好。

综合多标记评价 5 个指标, MUAE 方法的结果比其他 4 种方法和基准算法好, 而且在各组提取特征百分比情况下显示出好的稳定性。实验结果进一步证明, 自编码网络训练目标在各降维百分比情况下, 能保持甚至提取好的数据特征。

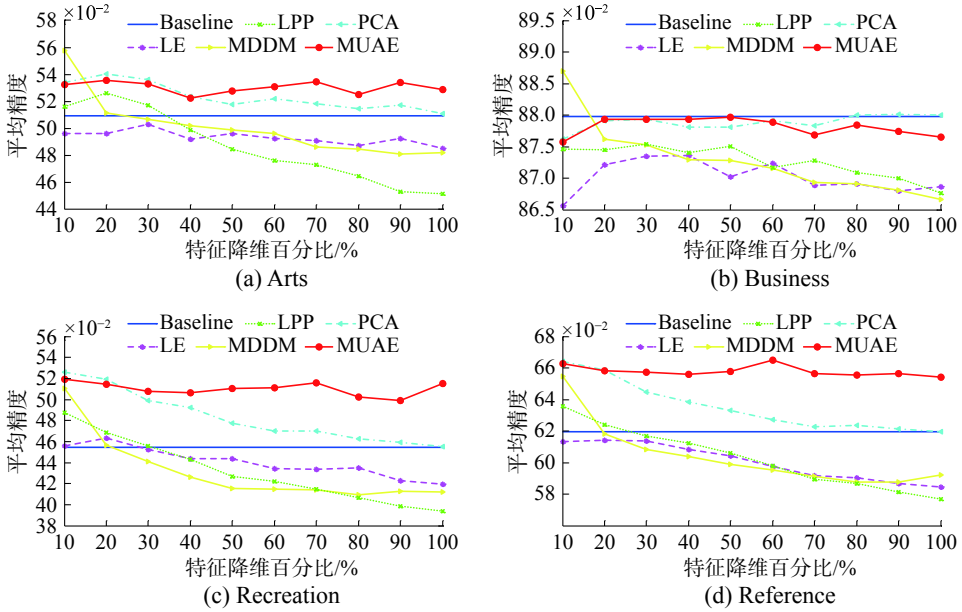


图 3　平均精度随特征降维百分比变化关系

Fig. 3　Relationship between average precision and percentage of features

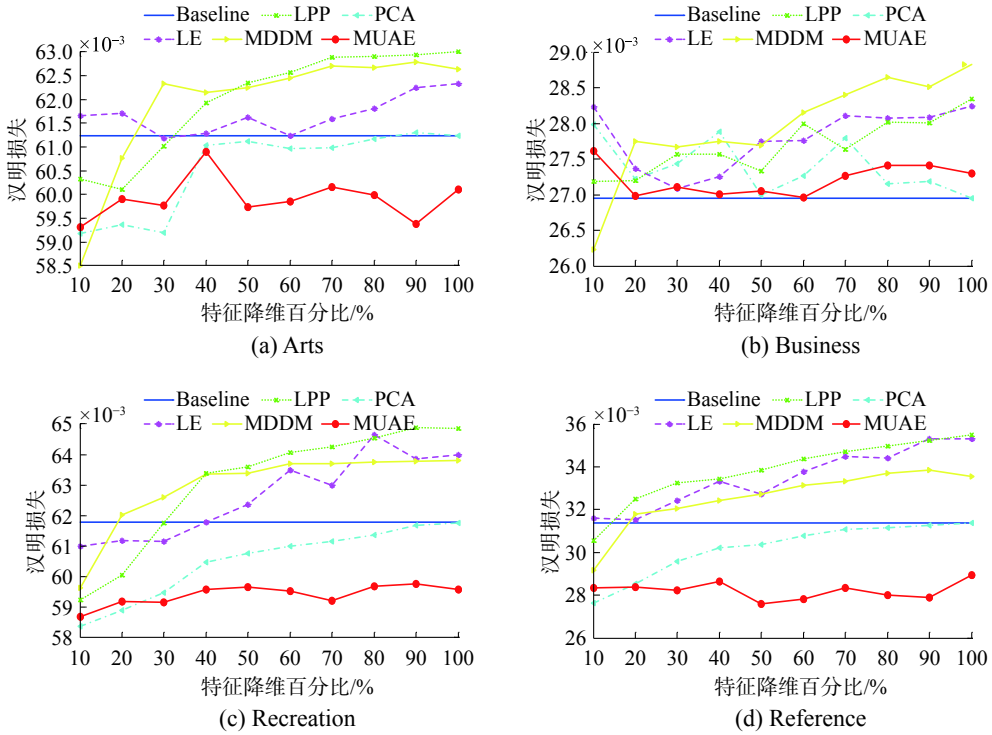
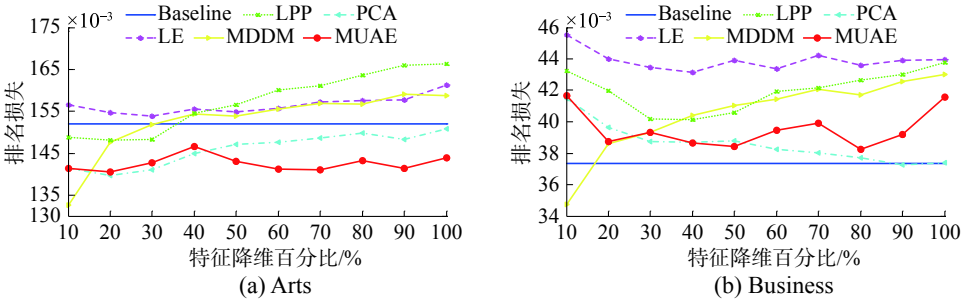


图 4　汉明损失随特征降维百分比变化关系

Fig. 4　Relationship between Hamming loss and percentage of features



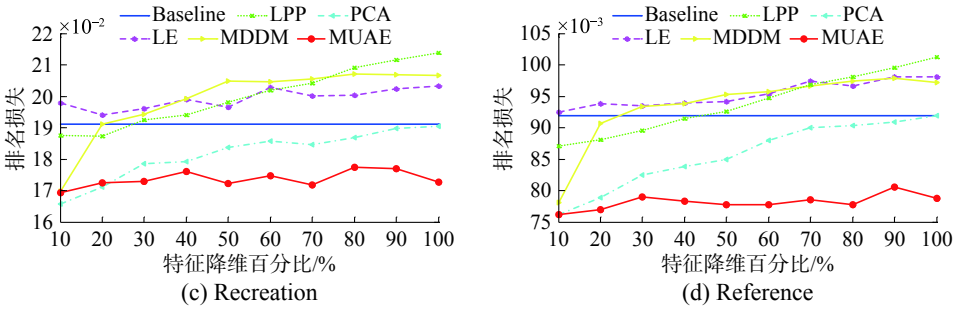


图 5　排名损失随特征降维百分比变化关系

Fig. 5　Relationship between ranking loss and percentage of features

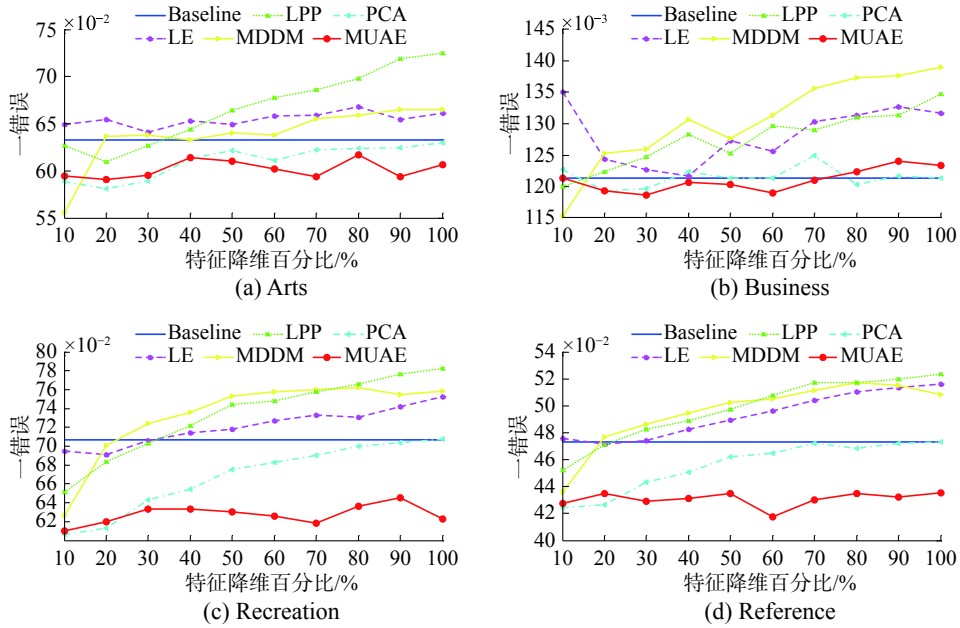


图 6　一错误随特征降维百分比变化关系

Fig. 6　Relationship between one-error and percentage of features

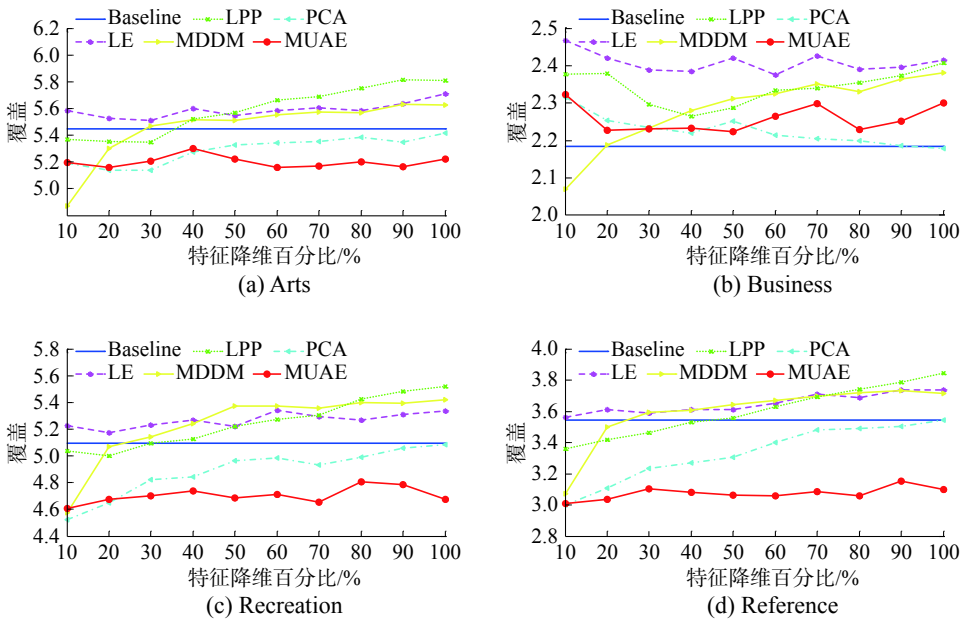


图 7　覆盖随特征降维百分比变化关系

Fig. 7　Relationship between voverage and percentage of features

5 结束语

针对多标记学习的数据降维问题提出自编码网络维数约简方法，用无监督方法处理有监督的多标记学习降维问题。通过实验验证了所构建自编码深度学习网络能自学习地提取多标记数据特征，降低数据维度，与其他无监督特征降维和多标记有监督降维方法相比，取得了较好的效果，在各百比降维的情况下，降维性能平稳性好。下一步工作，将使用变分自编码和降噪自编码网络对多标记和图像等数据进行无监督降维进行研究。

参考文献：

[1] ZHANG Minling, ZHOU Zhihua. A review on multi-Label learning algorithms[J]. IEEE transactions on knowledge and data engineering, 2014, 26(8): 1819–1837.

[2] TSOUMAKAS G, KATAKIS I. Multi-label classification: an overview[J]. International journal of data warehousing and mining, 2007, 3(3): 1–13.

[3] WU Fei, WANG Zhuhao, ZHANG Zhongfei, et al. Weakly semi-supervised deep learning for multi-label image annotation[J]. IEEE transactions on big data, 2015, 1(3): 109–122.

[4] LI Feng, MIAO Duoqian, PEDRYCZ W. Granular multi-label feature selection based on mutual information[J]. Pattern recognition, 2017, 67: 410–423.

[5] ZHANG Yin, ZHOU Zhihua. Multilabel dimensionality reduction via dependence maximization[J]. ACM transactions on knowledge discovery from data, 2010, 4(3): 1–21.

[6] 郭雨萌, 李国正. 一种多标记数据的过滤式特征选择框架[J]. 智能系统学报, 2014, 9(3): 292–297.
GUO Yumeng, LI Guozheng. A filtering framework fro the multi-label feature selection[J]. CAAI transactions on intelligent systems, 2014, 9(3): 292–297.

[7] JINDAL P, KUMAR D. A review on dimensionality reduction techniques[J]. International journal of computer applications, 2017, 173(2): 42–46.

[8] OMPRAKASH S, SUMIT S. A review on dimension reduction techniques in data mining[J]. Computer engineering and intelligent systems, 2018, 9(1): 7–14.

[9] YU Tingzhao, ZHANG Wensheng. Semisupervised multilabel learning with joint dimensionality reduction[J]. IEEE signal processing letters, 2016, 23(6): 795–799.

[10] YU Yanming, WANG Jun, TAN Qiaoyu, et al. Semi-supervised multi-label dimensionality reduction based on dependence maximization[J]. IEEE access, 2017, 5: 21927–21940.

[11] ZHANG Minling, ZHANG Kun. Multi-label learning by exploiting label dependency[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA, 2010: 999–1008.

[12] ZHANG Minling, ZHOU Zhihua. ML-KNN: a lazy learning approach to multi-label learning[J]. Pattern recognition, 2007, 40(7): 2038–2048.

[13] BALDI P. Autoencoders, unsupervised learning and deep architectures[C]//Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop. Washington, USA, 2011: 37–50.

[14] BOURLARD H, KAMP Y. Auto-association by multilayer perceptrons and singular value decomposition[J]. Biological cybernetics, 1988, 59(4/5): 291–294.

[15] 刘帅师, 程曦, 郭文燕, 等. 深度学习方法研究新进展[J]. 智能系统学报, 2016, 11(5): 567–577.
LIU Shuaishi, CHENG Xi, GUO Wenyan, et al. Progress report on new research in deep learning[J]. CAAI transactions on intelligent systems, 2016, 11(5): 567–577.

[16] VINCENT P, LAROCHELLE H, LAJOIE I, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion[J]. Journal of machine learning research, 2010, 11(12): 3371–3408.

[17] YU Ying, WANG Yinglong. Feature selection for multi-label learning using mutual information and GA[M]// MIAO D, PEDRYCZ W, ŚLĘZAK D, et al. Rough Sets and Knowledge Technology. Cham: Springer, 2014: 454–463.

[18] 余鹰. 多标记学习研究综述[J]. 计算机工程与应用, 2015, 51(17): 20–27.
YU Ying. Survey on multi-label learning[J]. Computer engineering and applications, 2015, 51(17): 20–27.

[19] 段洁, 胡清华, 张灵均, 等. 基于邻域粗糙集的多标记分类特征选择算法[J]. 计算机研究与发展, 2015, 52(1): 56–65.
DUAN Jie, HU Qinghua, ZHANG Lingjun, et al. Feature selection for multi-label classification based on neighborhood rough sets[J]. Journal of computer research and development, 2015, 52(1): 56–65.

[20] DOQUIRE G, VERLEYSEN M. Feature selection for multi-label classification problems[C]//Proceedings of the 11th International Conference on Artificial Neural Net-Works Conference on Advances in Computational Intelligence. Torremolinos-Mÿlaga, Spain, 2011: 9–16.

[21] LAMDA. Data & Code[EB/OL]. Nanjing: LAMDA, 2016[2018.03.20]. http://lamda.nju.edu.cn/Data.ashx.

[22] WOLD S, ESBENSEN K, GELADI P. Principal compon-

ent analysis[J]. Chemometrics and intelligent laboratory systems, 1987, 2(1): 27–52.

[23] HE Xiaofei. Locality preserving projections[M]. IL, USA: University of Chicago, 2005: 186–197.

[24] BELKIN M, NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural computation, 2003, 15(6): 1373–1396.

[25] SCHAPIRE R E, SINGER Y. Boos texter: a boosting-based systemfor text categorization[J]. Machine learning-special issue on information retrieval, 2000, 39(2/3): 135–168.

[26] TSOUMAKAS G, VLAHAVAS I. Random k-labelsets: an ensemble method for multilabel classification [M]//KOK J N, KORONACKI J, MANTARAS R L, et al. Machine Learning: ECML 2007. Berlin Heidelberg: Springer, 2007: 406–417.

作者简介：

杨文元，男，1967 年生，副教授，博士，主要研究方向为机器学习、多标记学习、模式识别、计算机视觉。发表学术论文 20 余篇。

第二届大数据和智能计算国际会议（ICBDSC 2019）
2019 2nd International Conference on Big Data and Smart Computing (ICBDSC 2019)

The Organizing Committee is pleased to announce that 2019 2nd International Conference on Big Data and Smart Computing (ICBDSC 2019) will take place in Bali, Indonesia on January 10–13, 2019 as a workshop of ICSIM 2019.

The conference is addressed to academics, researchers and professionals with a particular interest related to the conference topic. It brings together academics, researchers and professionals in the field of Big Data and Smart Computing making the conference a perfect platform to share experience, foster collaborations across industry and academia, and evaluate emerging technologies across the globe.

ICBDSC 2019 is co-organized by Universitas Pendidikan Ganesha (Undiksha), Indonesia and Kalbis Institute, Indonesia.

As a workshop of ICSIM 2019, accepted papers of ICBDSC 2019 will be collected in ICSIM 2019 conference proceedings, which will be submitted for indexing in Ei Compendex, Scopus, Thomson Reuters Conference Proceedings Citation Index etc. major data bases.