文章编号：2095-2163(2019)02-0042-05 中图分类号：TP183 文献标志码：A

基于稀疏自动编码器算法的 HBV 再激活分类预测模型

赵咏旺[1]，刘毅慧[1]，黄 伟[2]
（1 齐鲁工业大学(山东省科学院) 信息学院，济南 250353；2 山东省肿瘤医院 放疗病区，济南 250117）

摘 要：原发性肝癌患者在接受精确放疗后易引起乙型肝炎病毒(HBV)再激活。本文的研究目的就是根据已有的患者临床数据，建立分类预测模型来及时做出预测防护，从而在一定程度上降低 HBV 再激活的可能性。实验结果表明通过稀疏自动编码器特征提取的方法可以有效降低数据维度，提高预测准确度。Softmax 分类器对一层隐含层的稀疏自动编码器分类性能最优。5 折交叉验证下，平均准确率为 72.22%。SVM 分类器对二层隐含层的稀疏自动编码器分类性能最优。10 折交叉验证下，平均准确率为 78.52%。

关键词：原发性肝癌(PLC)；稀疏自动编码器(SAE)；特征提取；Softmax；SVM

HBV reactivation classification prediction model based on sparse autoencoder algorithm

ZHAO Yongwang [1], LIU Yihui [1], HUANG Wei [2]
(1 School of Information, Qilu University of Technology(Shandong Academy of Sciences), Jinan 250353, China;
2 Department of Radiation Oncology, Shandong Cancer Hospital, Jinan 250117, China)

【Abstract】 Patients with Primary Liver Cancer are susceptible to reactivation of Hepatitis B Virus (HBV) after receiving precise radiotherapy. The purpose of this study is to establish a classification prediction model based on existing patient clinical data to make predictive protection in time, thus reducing the possibility of HBV reactivation to a certain extent. The experimental results show that the method of feature extraction by sparse autoencoder can effectively reduce the data dimension and improve the prediction accuracy. The Softmax classifier has the best classification performance for the sparse autoencoder with one-hidden layer. Under the 50% cross-validation, the average accuracy is 72.22%. The SVM classifier has the best classification performance for the sparse autoencoder with two-hidden layers. Under the 10-fold cross-validation, the average accuracy is 78.52%.

【Key words】 Primary Liver Cancer (PLC); Sparse Automatic Encoder (SAE); feature extraction; Softmax; SVM

0 引 言

近年来，对原发性肝癌(Primary Liver Carcinoma，PLC)的诊治及术后已引发了研究学界的高度关注。目前，对中晚期原发性肝癌患者采用精确放疗后也会导致乙型肝炎病毒(Hepatitis B Virus，HBV)再激活[1]。因此，有效降低 HBV 感染率、发病率及死亡率至关重要。

当下研究中，文献[2]中接受 HSCT 的患者在单变量因子分析中检测到 HBV DNA 水平、年龄、HbsAg、HbcAb 与 HBV 再激活有关。在多变量分析中，年龄、HBsAg 是 HBV 再激活的危险因素。Ji 等人[3]通过应用 Kaplan-Meier 检验及 Cox 回归模型分析影响 HBV-PLC 患者生存期的因素。实验结果表明，Child-Pugh 分级、肿瘤转移、年龄、抗病毒治疗、血清 HBV DNA 水平和肿瘤治疗方式是影响患者生存期的重要因素。Han 等人[4]指出 HBV 再激活与肿瘤直径大小及是否术前规范抗病毒治疗等因素有关。Wang[5]通过研究比较基线特征差异筛选出 HBV 再激活的可能危险因素，结果显示性别、年龄等指标无明显差异，肝功能 Child-pugh 分级可能是 HBV 再激活的危险因素。Huang 等人[6]在 69 例原发性肝癌患者接受精确放疗后致使乙型肝炎病毒再激活研究中发现基线血清 HBV DNA 水平和放疗剂量是 HBV 病毒再激活的独立危险因素。Wu 等人[7]在以前发现的危险因素的基础上又研发建立了 RBF 神经网络模型，识别率提高到 80%。随后通过遗传算法发现 HBVDNA 水平、肿瘤分期 TNM、Child-Pugh、外放边界、V45 和全肝最大剂量是乙肝病毒再激活的危险因素[8]。Wang 等人[9-10]又分别

基金项目：国家自然科学基金(81402538，61375013)；山东省自然科学基金(ZR2013FM020)。

作者简介：赵咏旺(1992-)，男，硕士研究生，主要研究方向：人工智能、生物医学；刘毅慧(1965-)，女，博士，教授，主要研究方向：生物计算、智能信息处理；黄 伟(1979-)，男，博士，副主任医师，主要研究方向：肿瘤精确放射治疗的临床与基础研究。

通讯作者：刘毅慧 Email：yxl@sdili.edu.cn； 黄 伟 Email：alvinbird@163.com

收稿日期：2018-12-20

利用随机森林、小波变换、顺序前向、顺序后向等一系列特征选择方法使得 HBV 再激活的分类预测精度进一步提高。

特征提取就是指利用已有特征计算出一个抽象程度更高的特征集的过程。而稀疏自动编码器(Sparse Auto-Encoder)于 2007 年由 Bengio 提出，这就是一个典型 3 层结构神经网络，可以对特定数据进行关键特征提取[11]。其中，输入层和隐藏层之间的信号传递是编码过程，隐藏层与输出层之间的信号传递是解码过程，通过对输入数据的重构一方面可以检验自动编码器算法的学习效果，另一方面也可以对复杂的特征数据集进行降维，降维后的数据集表示输入数据最重要的因素[12]。本文通过采用稀疏自动编码器对山东省肿瘤医院提供的 90 例原发性肝癌患者精确放疗后影响 HBV 再激活的特征进行降维，并分别利用 SVM 及 Softmax 分类器对新的样本空间进行分类预测。对此拟展开研究论述如下。

1　数据来源

研究中，将山东省肿瘤医院的 90 例经过精确放疗后原发性肝癌患者的临床资料作为研究样本，每个样本包含有 28 个特征，组成 90 * 28 维大小的数据集，详情见表 1。

表 1　特征编号及分别对应的医学名称

Tab. 1　Feature numbers and the corresponding medical name

特征编号	医学参数
1	性别
2	年龄
3	KPS 评分
4	HbeAg
5	门脉癌栓有无
6	肿瘤分期 TNM
7	Child-Pugh
8	甲胎蛋白 AFP
9	HBV DNA 水平
10	放疗总剂量
11	等效生物剂量
12	放疗次数
13	放疗前 TACE
14	分割方式
15	GTV 体积(gross tumor volume)
16	PTV 体积(planning target volume)
17	外放边界
18	V5
19	V10
20	V15
21	V20
22	V25
23	V30
24	V35
25	V40
26	V45
27	全肝最大剂量
28	全肝平均剂量

2　算法研究与实现

2.1　稀疏自动编码器

稀疏自动编码器是人工神经网络的一种特殊学习模型。该模型输入输出是相同的，通过训练调整参数，使得输入的样本经过编码、以及解码变换后尽可能复现原来特征，具有良好特征提取能力[13]。本文利用稀疏自动编码器对原发性肝癌临床患者数据集进行特征提取，其中稀疏自动编码器隐含层分别取一到二层，其网络设计结构如图 1 所示。假设 F 和 G 分别表示编码和解码函数，可将其分别写作如下数学形式：

$$F(x) = s_l(\boldsymbol{w}_1 * x + p), \tag{1}$$

$$G(x) = p_l(\boldsymbol{w}_2 * h + q), \tag{2}$$

其中，s_l 为编码器激活函数，p_l 为解码器激活函数，本文研究中，分别选取 *satlin* 函数为编码器激活函数，*purelin* 函数为解码器激活函数，权值矩阵 $\boldsymbol{w}_1$，$\boldsymbol{w}_2$ 互为转置。

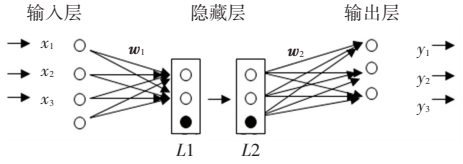

图 1　二层自动编码器结构示意图

Fig. 1　Two-layer automatic encoder structure diagram

研究中，p_i 表示第 $j(j = 1,2)$ 个隐藏层上的第 i 号神经元在训练集 $S = \{x^{(i)}\}_{j=1}^{N}$ 上的平均激活度。要求 $p_i = p(i = 1,2,3,\cdots,m)$，保证隐藏层上每个神经元都满足稀疏性限制。$p_i$ 的数学表述见如下：

$$p_i = \frac{1}{N} h_i(x^{(i)}), \tag{3}$$

其中，$h_i(n)$ 表示稀疏自动编码器隐含层数量为 $n(n = 2)$ 的第 i 号神经元的激活度[14-15]。本文中，$h(j = 1) = 20$，即第一层神经元节点数取值 20。$h(j = 2) = 15$，第二层神经元节点数取值 15。p 是稀疏性参数，$p(j = 1) = p(j = 2) = 0.015$，一层二层神经元稀疏性参数都是 0.015。正则化系数 $r(j = 1) = r(j = 2) = 2$，一层二层神经元稀疏性参数都是 2。*KL* 散度函数如下：

$$KL(p_i \| p) = p * \ln(\frac{p}{p_i}) + (1 - p) * \ln(\frac{1 - p}{1 - p_i}), \tag{4}$$

如果将 *KL* 函数加入到稀疏自动编码器的损失函数中，那么稀疏自动编码器的损失函数可表示为：

$$J_{SAE}(\theta) = \sum_{x \in S} L(x, G(F(x))) + \beta \sum_{i=1}^{m} KL(p_i \| p). \tag{5}$$

其中，β 为控制稀疏性惩罚的权重系数。当 $KL(p_i \| p) = 0$，也就是 $p_i = p$，此时就可以得到最小损失函数。

稀疏自动编码器是一个特征提取器，本实验要实现分类功能还要添加一个分类器，如贝叶斯分类器(Bayesian classifier)，支持向量机(Support Vector Machine)等[16]。本文主要采用的是 SVM 分类器，下面会将 Softmax 分类器的实验结果与 SVM 分类器的实验结果做出对比。

2.2 主成分分析

研究可知，作为特征提取的另一种方法，主成分分析算法(Principal Component Analysis)也可以对本实验数据集中的特征进行降维。PCA 特征降维的过程就是首先求得这个样本的协方差矩阵，并求出这个协方差矩阵的特征值与特征向量[17]。然后根据特征值大小，选取前三个特征值所对应的特征向量构成特征矩阵。最后用原始样本矩阵与得到的特征矩阵做积运算，会得到一个降维之后的新样本矩阵。为了保证原始测试样本同样能够映射到这个空间中进行表示，就需要将测试样本与之前对训练样本降维过程中得到的特征矩阵重新做积处理，便可以得到一个新的测试样本矩阵[18]。特征提取后的主要成分以及贡献率如图 2 所示。此后，研究中会针对不同主成分个数进行比较分析。

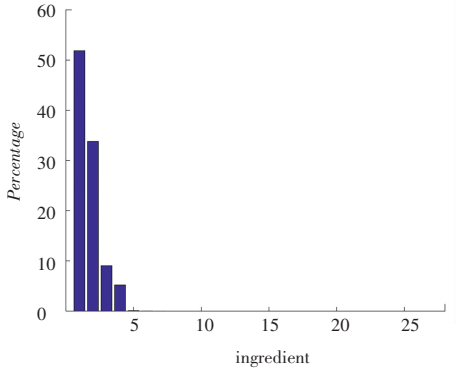

图 2 主成分贡献率排序

Fig. 2 Ordering of principal component contribution rates

2.3 分类器选择

本文主要选取 SVM 分类器进行最终分类处理，并用 Softmax 进行对照实验。作为一个二分类模型，SVM 的分类思想是给定一个包含正例和反例的样本集合，寻找一个超平面对样本根据正例和反例进行分割。其研究旨在使得分开的 2 个类别具有最大间隔，这样一来，分类才具有更高可信度以及更好的泛化能力[19]。假设超平面为 $\boldsymbol{w}x + b = 0$；样本点到超平面距离为：

$$\frac{t_i \cdot t(x_i)}{\| \boldsymbol{w} \|} = \frac{t_i \cdot (w^p \cdot \theta(x_i) + b)}{\| \boldsymbol{w} \|}, \tag{6}$$

首先，构造并求解约束最优化问题，研究推得数学运算公式如下：

$$\min(a): \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} a_i a_j t_i t_j (\theta(x_i) \cdot \theta(x_j)) - \sum_{i=1}^{p} a_i, \tag{7}$$

求得最优解 a^*，然后将用到公式(8)进行运算：

$$w^* = \sum_{i=1}^{p} a_i^* t_i \theta(x_i), \; b^* = t_i - \sum_{i=1}^{p} a_i^* t_i (\theta(x_i) \cdot \theta(x_j)), \tag{8}$$

最后求得分类决策函数，具体公式如下：

$$f(x) = sign(w^* \theta(x) + b^*). \tag{9}$$

文中，选取的是线性内核函数 $G(x_j, x_k) = x_j' x_k$。惩罚因子 c 是误差容忍系数。当 c 设置一个较大值时，表示要求的分类精度很高，分错的点会很少。当 c 设置一个较小值时，表示可能容忍一定的错误，分错的点可能就很多。由于本文的样本数据中乙型肝炎病毒再激活的类标签数量较少，在这里就必须保证其分类正确率。本文中，c 的取值是 2。

Softmax 模型是 logistic 回归模型在多分类问题上的推广，当 Softmax 是一个二分类处理器时就会成为一个 Logistic 分类[20]。在本文 Softmax 分类层中，选取交叉熵函数作为损失函数，函数公式如下：

$$L = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{k} \boldsymbol{t}_{ij} \ln y_{ij} + (1 - t_{ij}) \ln(1 - y_{ij}). \tag{10}$$

其中，n 是训练样本数量 $k = 2$，即二分类问题；t_{ij} 是目标矩阵 $\boldsymbol{t}$ 的 i 行 j 列的元素；y_{ij} 是输入向量为 $\boldsymbol{x}_j$ 时自动编码器的第 i 个输出。

3 实验结果及分析

3.1 分类性能度量

本文主要采用 3 个分类性能指标，分别是准确性、特异性、灵敏性。其中，准确性是指分类的正确预测值占样本实际值的比重。特异性是将实际无病的人正确判定为真阴性的比例。灵敏性是将实际有病的人正确判定为真阳性的比例。

3.2 实验结果

实验分别采用 3 折、5 折、10 折交叉验证，选取

每一个分类性能度量标准的平均值作为最终数据，测试实验结果见表 2～表 4。

表 2　Softmax 对不同隐含层数 SAE 对比实验结果

Tab. 2　Experimental results with Softmax classifier for different hidden layers SAE

隐含层数	*K* 折	预测精度	特异性	灵敏性
原始数据	3	0.697 8	0.875 4	0.114 3
	5	0.701 9	0.878 6	0.083 3
	10	0.711 1	0.881 0	0.116 7
1	3	0.707 8	0.876 8	0.152 4
	5	0.722 2	0.878 6	0.175 0
	10	0.718 5	0.871 4	0.183 3
2	3	0.691 1	0.856 5	0.147 6
	5	0.690 7	0.816 7	0.250 0
	10	0.700 0	0.871 4	0.100 0

表 3　SVM 对不同隐含层数 SAE 对比实验结果

Tab. 3　Experimental results with SVM classifier for different hidden layers SAE

隐含层数	*K* 折	预测精度	特异性	灵敏性
原始数据	3	0.703 2	0.878 3	0.142 9
	5	0.707 4	0.878 6	0.108 3
	10	0.714 8	0.876 2	0.150 0
1	3	0.740 0	0.943 5	0.071 4
	5	0.751 9	0.940 5	0.091 7
	10	0.729 6	0.914 3	0.083 3
2	3	0.763 3	0.988 4	0.033 8
	5	0.777 5	0.971 4	0.100 0
	10	0.785 2	0.995 2	0.050 0

表 4　PCA 不同主成分个数下分类预测实验结果

Tab. 4　Experimental results of classification prediction under different number of principal components

主成分个数	贡献率/%	预测精度	特异性	灵敏性
1	51.86	0.666 7	0.842 9	0.050 0
2	85.65	0.720 5	0.900 0	0.050 0
3	94.68	0.750 0	0.928 6	0.024 3
4	99.88	0.759 2	0.928 6	0.126 7

表 2 是 Softmax 分类器分别对原始数据集、一层、二层 SAE 提取的特征进行分类预测的结果。由表 2 可知，对原始数据集的预测准确率要低于对一层、二层 SAE 所提取特征数据的预测准确率。3 折、5 折、10 折交叉验证下，Softmax 分类器在一层 SAE 下的预测结果都要高于在二层 SAE 下的预测结果。其中，5 折交叉验证下，SVM 对一层 SAE 提取特征数据集的识别率可达 72.22%，比二层 SAE 的识别率提高了近 3.2 个百分点。从灵敏性结果来看，3 折、5 折、10 折交叉验证下，对一层、二层 SAE 所提取特征数据的灵敏度表现都要优于对原始数据集的灵敏度表现。

表 3 是 SVM 分类器分别对原始数据集、一层、二层 SAE 提取的特征进行分类预测的结果。由表 3 可知，对一层、二层 SAE 提取的特征进行分类预测的结果要明显优于未经过特征提取的原始数据预测结果。这一点与 Softmax 分类器的表现是一致的。而灵敏度的表现却截然相反，SVM 对原始数据的分类灵敏度要略高于对 SAE 特征提取数据的分类灵敏度。在 3 折、5 折、10 折交叉验证下，SVM 分类器在二层 SAE 下的预测结果都要高于在一层 SAE 下的预测结果。其中，10 折交叉验证下，SVM 对二层 SAE 所提取特征数据集的预测准确率可达 78.52%，比原始数据集的预测准确率提高了近 7.0 个百分点，比一层 SAE 的预测准确率提高了近 5.6 个百分点。

综合来看，虽然 SVM 及 Softmax 对原始数据的分类预测表现相差不大，但从经过一层 SAE 或者二层 SAE 提取特征之后的预测准确率角度来看，SVM 的表现要更加优越。经过一层 SAE 提取特征之后，在 3 折交叉验证下，SVM 预测精度高于 Softmax 预测精度近 3.2 个百分点。在 5 折交叉验证下，SVM 预测精度高于 Softmax 预测精度近 3.0 个百分点。经过二层 SAE 提取特征之后，在 3 折交叉验证下，SVM 预测精度高于 Softmax 预测精度近 7.2 个百分点。在 5 折交叉验证下，SVM 预测精度高于 Softmax 预测精度近 8.7 个百分点。在 10 折交叉验证下，SVM 预测精度高于 Softmax 预测精度近 8.5 个百分点。

表 4 是 PCA 在 5 折交叉验证下特征提取并建立 SVM 模型分类预测的结果。由表 4 可以看出，在不同主成分个数下的预测精度是有区别的。随着主成分个数的增加，分类预测精度也是呈递增趋势。4 个主成分的贡献率是 99.88%，而除去前 4 个主成分之外的其它成分贡献率都在 0.01% 以下，可以看作是冗余信息，所以主成分个数最多取到 4 个，此时预测精度是 75.92%。灵敏性也是最高的，相比 3 个主成分时提高了近 10.2 个百分点。同样在 5 折交叉验证下，SVM 在二层 SAE 提取特征下的预测精度是 77.75%，相比 PCA 在 4 个主成分下的预测精度提高了近 1.8 个百分点。由此可见，在对本实验的样本数据进行重要成分提取压缩的过程中，SAE 的效果要优于 PCA。

4 结束语

原发性肝癌患者在精确放疗后乙型肝炎病毒再激活是一种常见并发症,及时的预测防护能降低发病率、死亡率。影响原发性肝癌患者发生 HBV 再激活的危险因素有很多,通过构建二层学习的稀疏自动编码器相对主成分分析算法更能有效地对原发性肝癌患者临床数据中的重要成分进行提取。而 SVM 分类器有效提高了 HBV 再激活的分类预测准确性,并且对二层稀疏自动编码器分类性能堪称最优,10 折交叉验证下,平均准确率达 78.52%。

参考文献

[1] LAVANCHY D. Hepatitis B virus epidemiology, disease burden, treatment, and current and emerging prevention and control measures[J]. Journal of Viral Hepatitis, 2004,11(2):97-107.

[2] JUN C H, KIM B S, OAK C Y, et al. HBV reactivation risk factors in patients with chronic HBV infection with low replicative state and resolved HBV infection undergoing hematopoietic stem cell transplantation in Korea[J]. Hepatology International, 2017, 11(1):87-95.

[3] JI Wei, WANG Wei, LYU Jinhan, et al. Risk factors influencing survival of patients with hepatitis B virus-related primary liver cancer[J]. Journal of Practical Hepatology, 2015, 18(6):638-642.

[4] HAN Juqiang, REN Yongqiang, LI Guo´an. A. Study on reactivation HBV and related influencing factors after minimally invasive interventional therapy for primary hepatic cancer[J]. Chinese Frontiers of Medicine: Electronic Edition, 2014,6(3):27-30.

[5] WANG Mengsen. Study on reactivation of hepatitis B virus by three dimensional conformal radiotherapy for primary hepatic carcinoma[D]. Jinan: University of Jinan,2014.

[6] HUANG Wei, ZHANG Wei, FAN Min, et al. Risk factors for hepatitis B virus reactivation after conformal radiotherapy in patients with hepatocellular carcinoma[J]. Cancer Science, 2014, 105(6):697-703.

[7] WU Guanpeng, LIU Yihui, WANG Shuai,et al. HBV reactivation classification prediction model based on feature selection of genetic algorithm[J]. Chinese Journal of Bioinformatics, 2016,14(4): 243-248.

[8] WU Guanpeng, LIU Yihui, WANG Shuai, et al . The classification prognosis models of hepatitis b virus reactivation based on Bayes and support vector machine after feature extraction of genetic algorithm[C]//2016 12th International Conference on Natural Computation and 13th Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). Changsha,China:IEEE, 2016:572-577.

[9] WANG Huina, HUANG Wei, LIU Yihui. Random forest and Bayesian prediction for Hepatitis B virus reactivation[C]// International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). Guilin, China: IEEE, 2017:2060-2064.

[10] WANG Huina, LIU Yihui, HUANG Wei. The application of feature selection in hepatitis B virus reactivation[C]//2017 IEEE International Conference On Big Data Analysis. Beijing, China: IEEE,2017:893-896.

[11] LIN Xinyu, ZHU Ce, ZHANG Qian, et al. 3D keypoint detection based on deep neural network with sparse auto-encoder[J]. arXiv Preprint arXiv: 1605.00129,2016.

[12] LI Hang, WANG Haozheng, YANG Zhenglu, et al. Variation auto-encoder based network representation learning for classification[C]// ACL 2017-Student Research Workshop. Vancouver,Canada:ACL, 2017:56-61.

[13] JALALVAND S, FALAVIGNA D. Stacked auto-encoder for ASR error detection and word error rate prediction[C]// ISCA Interspeech2015. Dresden, Germany:ISCA,2015:2142-2146.

[14] FERNANDES S, SOUSA R G, SOCODATO R, et al. Stacked denoising auto-encoders for the automatic recognition of microglial cells´ state[C]// Esann2016. Bruges, Belgium:[s.n.],2016:1-7.

[15] 刘芳, 路丽霞, 黄光伟,等. 基于稀疏自动编码器和支持向量机的图像分类方法:中国, CN106529574A[P]. 2017-03-22.

[16] 邓俊锋, 张晓龙. 基于自动编码器组合的深度学习优化方法[J]. 计算机应用, 2016,36(3):697-702.

[17] CHEN Xianmin, ZHANG Chaoyang, ZHOU Zhaoxian. Improve recognition performance by hybridizing principal component analysis (PCA) and elastic bunch graph matching (EBGM)[C]//2014 IEEE Symposium on Computational Intelligence for Multimedia, Signal and Vision Processing (CIMSIVP). Orlando, FL, USA:IEEE, 2014:1-5.

[18] KEPCEOGLU A, GÜNDOGDU Y, LEDINGHAM K W D, et al. Identification of the isomers using principal component analysis (PCA) method[C]// 9th International Physics Conference of the Balkan Physical Union. Istanbul: AIP Publishing LLC, 2016: 060004(1-4).

[19] GAO Yan, ZHOU Chenghu, SU Fenzhen. Study on SVM classifications with multi-features of OLI images[J]. Engineering of Surveying & Mapping,2014, 23(6):1-5,10.

[20] GIMPEL K, SMITH N A. Softmax-margin CRFs: Training log-linear models with cost functions[C]//Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics. Los Angeles, California, USA:dblp,2010:1-4.