一种解决英语动名词搭配错误的模型

杜一民 吴桂兴 吴 敏
(中国科学技术大学软件学院 苏州 215123)

摘 要 英语学习者易犯动名词搭配错误。通过分析CLEC中的动名词搭配错误,提出一种纠正中国学习者的动名词搭配错误的模型。首先构建了一个动名词搭配库,接着提出了一种度量搭配间的相似度的方法,通过计算目标搭配和搭配库的相似度得到粗略的相似搭配集,使用分类的方法过滤掉相似搭配集中不能将测试句划为正确类的搭配,得到候选结果集,最后使用语言模型对候选结果集打分排序,得到最终的纠正建议。在使用BNC构造的测试集上,这种综合相似性推理和上下文特征的方法对动名词搭配纠错具有显著效果。

关键词 搭配错误,分类,语言模型
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.7.041

Model to Solve English Verb-Noun Collocation Errors

DU Yi-min WU Gui-xing WU Min
(School of Software Engineering, University of Science and Technology of China, Suzhou 215123, China)

Abstract English learners often make mistakes about verb-noun collocations. Through the analysis of verb-noun collocation errors in CLEC, a model was proposed to correct these mistakes made by Chinese learners in this paper. A library on verb-noun collocations was first built, a method was then put forward to measure the similarity between the collocations, and similarity between the target and the library was calculated to get a coarse similar-collocation set. After that, a classification was applied to filer out incorrect collocations, and the final candidate set would be ranked by a language model to obtain the correction suggestions. In the test data constructed by the BNC corpus, this method which combines similarity inference and contextual features has a significant effect on verb-noun collocation errors' correction.

Keywords Collocation error, Classification, Language model

搭配错误是英语语法错误中的一种,解决包含搭配错误在内的英语语法检错纠错问题是自然语言处理领域的重要任务之一。文献[1]指出,据"中国学习者语料库"的统计,在所有的言语失误中,搭配错误在言语失误频率表中位居第六。其中,动名词搭配错误出现的比例占第一位。因此,解决动名词搭配的检错纠错问题是语法错误纠正(GEC)问题的重要环节。

目前,动名词搭配检错纠错的主流方法主要有以下几种:基于错误实例匹配的方法、基于概率模型的方法和基于机器学习的方法。传统的基于错误实例匹配的方式采用人工标记常见搭配错误类型并存入数据库,这种方法只能通过字符串匹配的方式识别错误并且只能提供预先存储的纠正建议,构建一个这样的数据库耗时耗力,并且纠错能力受制约于预存的纠正建议。文献[2]采用了一种基于概率的模型,这种方法将搭配本身的特征整合到一个概率模型中,通过综合这些特征对候选搭配集进行排名,这种方法充分考虑了动名词搭配本身的特点,但是没有利用到搭配词周围的上下文关系。文献[3]使用了一种基于最大熵的分类算法来进行动名词搭配的检错纠错,这种方法考虑了搭配词周围的局部上下文关系,但是该方法只对动名词搭配中的动词进行检错纠错,而将名词作为分类器的特征参与训练。

以上主流方法并没有充分分析语言学习者的错误行为和错误动机。文献[4]建议采用基于语料库的研究方法,通过对错误赋码与搭配词进行检索,提取动名词搭配错误,对搭配错误进行分类及诊断分析,以此解释动名词搭配错误产生的原因,并据此提出尝试性的教学建议。在对于动名词搭配错误的研究中,文献[5]对中国学习者英语语料库中大学英语四、六级两个子语料库中的动名词搭配错误进行分析,将动名词搭配错误分为以下几类:动词、名词、同义词使用错误,学习者"创造"搭配、超常搭配、直译造成的错误。

本文提出了一种解决英语中动名词搭配错误的方法,该方法将为常见动名词搭配构建一个搭配库,并为搭配库中的动名词搭配训练对应的分类器。在错误检测过程中,采用一种相似性计算的算法找到被检测搭配的相似搭配集。相似搭配集中每个搭配的分类器可以计算出该搭配是否符合待测试语句的语境,符合待测句语境的相似搭配将组成候选搭配集。一个经过训练的语言模型将把候选搭配集中的每个搭配替换到待测句中并计算得分,按照得分高低排序即得最终改正建

到稿日期:2015-05-15 返修日期:2015-08-04 本文受江苏省科技项目-基础研究计划(自然科学基金)面上研究项目(BK20141209)资助。
杜一民(1991-),男,硕士,主要研究方向为自然语言处理、信息检索,E-mail:sa613403@mail.ustc.edu.cn;吴桂兴(1972-),男,博士,讲师,主要研究方向为多媒体信号处理、嵌入式软件与数据挖掘等;吴 敏(1962-),男,教授,主要研究方向为教育软件工程、e-Learning、国外高等教育比较研究等。

议。在为模拟 CLEC 中动名词搭配错误行为而构造的测试集上的实验表明，这种综合考虑搭配错误行为、搭配相似性推理和搭配错误上下文特征而设计出的方法能非常有效地改正中国英语学者的动名词搭配错误。

1 模型设计

本文所设计的模型主要由 3 个部分组成：动名词搭配之间的相似性度量、相似搭配集的过滤、候选结果集的重排序，该模型执行的具体过程如下：

1)首先提取测试句子中的动名词搭配，然后用其与搭配库中的每一个搭配进行相似度计算，取相似性最高的 n 个搭配得到相似搭配集。

2)将一个搭配是否符合当前测试句的语境转化为测试句针对该搭配的分类问题，使用相似搭配集中每个搭配的分类器分别对测试句进行分类，把能将测试句分为正确类的搭配过滤出来组成候选结果集。

3)将候选结果集中的搭配分别替换回原测试句并使用语言模型计算每个句子的概率，通过对概率的排序确定最终的改正结果列表。

整个模型的系统流程如图 1 所示。

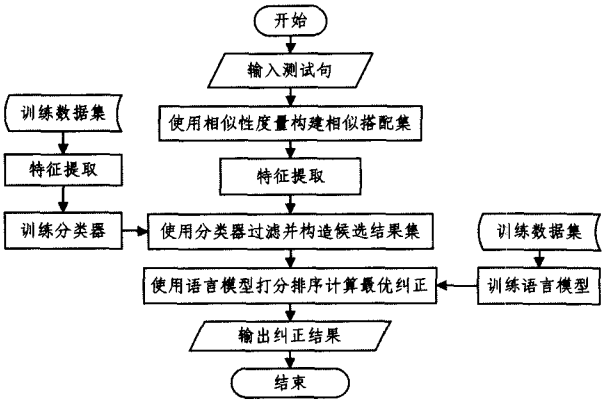

图 1 动名词搭配检错纠错系统流程

1.1 构建相似性度量

构建相似性度量的目的是能在搭配库中甄选出一个包含改正搭配的相似搭配集。通过对中国英语学习者语料库中的动名词搭配错误的分析可知[5]，动名词搭配错误的表现形式主要是动词、名词、同义词使用错误及直译错误等，所以错误搭配与其纠正结果应该具有较高的语义相似度。

本文在计算搭配之间的相似度时综合考虑动词之间和名词之间的相似度，在计算词之间的相似度时使用了 Jiang-Conrath 估计方法。Jiang-Conrath 估计是一种结合了基于语料库统计分析和基于词汇分类结构(WordNet)来计算词之间的语义相似度的方法，该方法综合考虑词语之间的结构信息和内容信息，优于其他基于词汇分类结构的算法，并接近人工判定词语间相似性精度的理论上界[6]。文献[6]给出了 Jiang-Conrath 估计方法的简单定义，式(1)计算所得为概念 a 和概念 b 之间的语义距离 $dist_{JCN}(a,b)$：

$$dist_{JCN}(a,b)=\max[IC(a)+IC(b)-2*IC(LCS(a,b))] \tag{1}$$

其中，$LCS(a,b)$代表概念 a 和 b 的最近公共祖先。IC 代表概念的信息内容，它的定义为：

$$IC(c)=\log^{-1}P(c) \tag{2}$$

其中，$P(c)$代表在大规模训练语料中概念 c 的概率。

概念 a 和概念 b 间的相似度定义为：

$$sim_{JCN}(a,b)=\frac{1}{dist_{JCN}(a,b)} \tag{3}$$

上述相似度的值越高表明两个概念之间的相似性越强。本文定义动名词搭配之间的语义相似度为：

$$sim_{JCN}(vn_1,vn_2)=sim_{JCN}(v_1,v_2)+sim_{JCN}(n_1,n_2) \tag{4}$$

由于在训练语料中 do，make，take 等虚化动词出现的频率较高[7]，导致其 IC 值过大，这些词之间的相似度会比较大，使得包含虚化动词的搭配之间的相似度明显大于包含虚化动词和包含普通动词的搭配之间的相似度，因此在实际的计算中，对于包含虚化动词的搭配间的相似度计算结果值，需要对其按比例缩小。

本文在计算词语之间的相似性度时使用 WordNet 的公共访问接口来对词汇分类树进行访问，并且使用大规模的英语语料计算出英语单词的信息内容(IC)。

在对测试句子进行检错时，使用该句子中的动名词搭配依次与搭配库中的动名词搭配计算语义相似度，并对语义相似度进行排名，取语义相似度最高的 n 个搭配组成相似搭配集。本文构造的搭配库的依据是对错误搭配中的动词、名词的近义词替换，因此使用上述计算方式理论上能有效地将相似搭配集从整个搭配库中甄选出来。

1.2 用分类器对相似搭配集进行过滤

上一节计算得到的相似搭配集是一个比较粗糙的结果集，该结果集并不能直接用来提供建议结果，需要将无关的搭配过滤掉以构建一个精简的候选结果集，对无关搭配的过滤需要考虑上下文特征，基于统计的方法可以充分利用上下文特征。文献[8]使用了 4 种统计学习方法来改正英语学习者的语法错误，结合具有相同特征的数据，经过对比表明感知机这类线性算法能够取得比朴素贝叶斯算法和另外两种语言模型更好的效果。

本文使用与感知机算法具有相同算法结构的被动主动算法(PA)[9]，文献[9]表明 PA 算法的效果优于感知机算法，PA 算法结合了感知机算法和 SVM 的优点，学习速度更快，更易实现。为了避免过度拟合，在训练权重向量时使用了平均策略，算法定义如下(算法 1 中 K 表示算法迭代的轮数；cw 表示多轮迭代中的权重向量和，用于计算多轮迭代的平均权重；输入的训练样本数据为 n 维特征向量，特征向量的生成规则见下文，输入的训练样本标签取值范围为$\{-1,1\}$，输出 w 为 n 维权重向量)。

算法 1 Passive-Aggressive 算法

输入：主动性参数 C>0，K，样本数 T
输出：w
初始化：cw=(0,…,0)
for k=0…K−1 do
  $w_1=(0,\cdots,0)$;
  for t=1…T−1 do
    选取样本：$x_t\in R^n$;
    预测样本的标签：$\hat{y}_t=\text{sign}(w_t * x_t)$
    获取样本的正确标签：$y_t\in\{-1,+1\}$
    计算损失精度 $\ell_t(w;(x,y))$;
    用式(5)更新权重向量得到 w_{t+1};
  end
  cw=cw+w_T;
end

w=cw / K;

参数更新策略使用如下公式：

$$w_{t+1}=w_t+\tau_t^* x_t y_t \tag{5}$$

式(5)中的更新参数 τ_t^* 定义如下：

$$\tau_t^*=\min(C,\tau_t) \tag{6}$$

算法中的 C 是用来控制更新权重的主动性程度的参数，$C>0$。其中，

$$\tau_t=\frac{\ell_t}{\|x_t\|^2} \tag{7}$$

式(7)中的 ℓ_t 为算法采用的损失函数，本算法使用的损失函数为 *hinge loss*，其定义为：

$$\ell_t(w;(x,y))=\begin{cases}0, & y_t(w_t\cdot x_t)\geqslant 1\\ 1-y_t(w_t\cdot x_t), & \text{otherwise}\end{cases} \tag{8}$$

式(8)中，$y_t(w_t\cdot x_t)$定义为计算第 t 个样本实例时的边际距离。

PA 算法使用一种侵略性的策略修改权重向量，以尽可能地满足对当前样本的约束。PA 算法是一种支持多分类的线性算法，本文使用二元分类算法来实现过滤器的功能，每个搭配的权重向量存储在一个以该搭配词组命名的文件中。

有效的特征选择会提高分类的精度，本文使用了动名词搭配的上下文特征作为分类的特征，采用了搭配上下文的 n 元组，参考了文献[3]的提取方法。文献[3]除了使用搭配上下文的 n 元组作为特征外，还使用了目标搭配中的名词作为特征，原因是其认为动名词搭配错误是由动词选择错误导致，故只对动名词搭配中的动词进行纠正。本文认为动名词搭配错误与动词和名词的使用都有关系，因此动词和名词都是本文的纠正目标，所以未使用搭配中的名词作为特征。

本文使用目标搭配词周围的一元组和二元组作为提取的上下文特征。在处理中，目标搭配周围的标点符号不会被计入 n 元组中，下面是一个例子。

例句：I will give you an example of why I have come to that conclusion.

以句中的 give example 为目标搭配提取出的 n 元组特征如下：

UniVL=will	UniVLL=I	BiVL=I will
UniVR=you	UniVRR=an	BiVR=you an
UniNL=an	UniNLL=you	BiNL=you an
UniNR=of	UniNRR=why	BiVR=of why
BiVI=will you	BiNI=an of	

其中，Uni 代表一元组，Bi 代表二元组，V 代表目标搭配的动词，N 代表目标搭配的名词，L 代表目标词左边第一个词，LL 代表目标词左边第二词；同理，R 和 RR 代表目标词右边第一个词和目标词右边第二个词，I 代表目标词在二元组的中间。

1.3 基于语言模型的候选结果集重排序

候选结果集需要经过排序才能形成最终的建议改正列表，对候选结果集中搭配的排序需要考虑整个测试句子的上下文特征，文献[10]表明语言模型适合解决此类问题。因此本文在此处使用了三元语言模型[11]，模型的定义如下：

三元语言模型由 V 和参数 $q(w|u,v)$ 组成，其中 V 为语言中所有单词的集合，$w\in V\cup\{\text{STOP}\}$，$u,v\in V\cup\{*\}$。STOP 定义为句子的结尾，* 定义为句子的开头。$q(w|u,v)$ 的含义是句子中的前两个单词为 (u,v) 的情况下下一个词是 w 的概率。

定义一个句子为 $x_1,x_2,\cdots,x_n$，其长度为 n，其中 x_n 恒为 STOP，假设 $x_0=x_1=*$，根据二阶马尔科夫模型可得，句子 $x_1,x_2,\cdots x_n$ 的概率为：

$$p(x_1\cdots x_n)=\prod_{i=1}^{n}q(x_i|x_{i-2},x_{i-1}) \tag{9}$$

对于概率 $q(w|u,v)$，使用最大似然估计计算：

$$q(w|u,v)=\frac{c(u,v,w)}{c(u,v)} \tag{10}$$

其中，$c(u,v,w)$表示训练语料中的(u,v,w)三元组的出现次数，$c(u,v)$表示训练语料中的(u,v)二元组出现的次数。由于在训练中，可能会出现 $c(u,v,w)=0$ 的情况，因此导致 $q(w|u,v)=0$，对于这种数据稀疏问题，需要使用参数平滑处理。

本文对语言模型的训练使用的数据是整个 BNC 语料库，三元语言模型的实现参照伯克利大学的 n-gram 语言模型[12]，在模型训练中使用了 Kneser-Ney 平滑方法[13]。

2 实验设计

本文主要研究中国英语学习者的动名词搭配错误的检错纠错，通过对 CLEC 语料库的分析，首先从中找出包含有动名词搭配错误(错误标签是 cc3)的句子，然后找出其中的错误动名词搭配，接着由两名研究人员分别根据上下文语境和牛津搭配词典的相似例句对该错误进行改正，然后进行对比和协商，最终取得一致的改正，本文使用其中的 90 组来进行实验。针对错误搭配以及它的改正，从牛津搭配词典中找到 4 至 6 个相似搭配，这些相似搭配将组成改正搭配的混淆集。把所有的改正搭配和相似搭配都存入搭配库，对于搭配库中的每一个搭配，从网络语料库中收集 50 句包含该搭配的英语句子存入以每一个搭配命名的文件中，作为每一个搭配的训练集。

本文使用整个 BNC 语料库来训练语言模型和单词的 IC 值。首先解析 BNC 语料库，然后将其按每行一句的方式存储到文本文件中作为训练语料。

测试集的构建方法：为每一个错误搭配所对应的改正搭配收集若干条句子，其中的一半句子不做修改作为正测试例，另一半句子注入该错误搭配作为测试的负测试例，将所有的测试句按行存入测试文件。

实验过程中所用的提取动名词搭配的方法使用了 Stanford Parser 中的依赖关系提取方法，使用该方法提取的依赖关系中的形如 dobj(give-3,example-6)的依赖关系即为所要提取的动名词搭配。除此之外，由于动词时态和名词的单复数在每个句子中都有所不同造成相同搭配具有多种不同形式，因此将实验中所有的动名词搭配中的动词和名词都做了还原词干的处理，此处使用 Stanford-CoreNlp 提供的方法完成了词干的还原。

3 实验结果分析

3.1 近似搭配计算结果分析

针对上文收集的 90 组错误搭配，用每一组错误搭配的改正搭配及其相似搭配共同组成一个动名词搭配库，本文构建的搭配库中共有 500 个搭配。针对测试句子中的动名词搭配，分别与搭配库中的搭配进行相似度计算，并统计相似度最大的前 n 个搭配中是否包含该错误搭配的改正。为了证明相似搭配计算的可扩展性，实验中又从 BNC 和牛津搭配词典中额外收集了 2000 个动名词搭配，构建了一个包含 2500 个搭配的搭配库，新增的 2000 个搭配中有部分与原搭配库中的

500 个搭配相似。针对不同的 n 值，分别对两个搭配库统计，计算结果中改正搭配的覆盖率。结果如表 1 所列。

表 1　最相似前 n 个搭配中改正搭配覆盖率(%)

n	5	10	15
覆盖率(500 个搭配)	74.5	85.4	90.1
覆盖率(2500 个搭配)	59.5	77.5	85.3

根据表 1 可知，当使用本文构建的小规模搭配库时，通过本文的相似搭配计算方法所得的前 n 个结果的改正搭配覆盖率比较大，而且随着 n 的增加，包含改正搭配的概率越大。当搭配库中搭配数目的规模变大时，只要适当地将 n 值变大，仍能保证较大的改正搭配覆盖率。

3.2　最终结果分析

在对模型的整体测试中，本文通过收集包含改正搭配的句子构建了一个测试集，然后人工地将错误搭配注入测试集中。对于每个改正搭配收集了 10 个句子，将其中的 5 句人工地注入其对应的错误搭配，对另外 5 句不作处理，测试集共有 900 个句子。

本文对测试结果的评判使用的是准确率、召回率和平均倒数排名(MRR)[14]。准确率表示被该模型正确改正了的句子数占总句子数的百分比，从总体上反映了一个句子能够被正确纠正的情况。召回率表示正确句子最终被标记为正确的比率，反映了正确测试句不被误改的比率。准确率和召回率都是针对建议结果列表的第一个结果来计算的。MRR 是一种搜索引擎的排序评价方法，考虑的是整个建议结果列表，反映了建议结果列表中是否包含正确改正以及正确改正是否在列表的前面，计算方法为：排名第一的结果和正确结果匹配，分数为 1；第二个结果和正确结果匹配，分数为 0.5；第 n 个结果和正确结果匹配，则分数为 $1/n$；没有结果与正确结果匹配，则分数为 0。最终得分为所有测试句计算结果的平均值。

针对 $n=5$，$n=10$，$n=15$，对测试数据集进行计算得到的准确率、召回率、MRR 值如表 2 所列。

表 2　不同相似列表长度下的准确率、召回率、MRR(%)

n	5	10	15
准确率	71.5	72.4	75.1
召回率	83.5	79.3	74.2
MRR	78.3	78.8	82.7

通过分析表 2 可知，提出的模型总体上具有比较高的准确率、召回率和 MRR 值。当取计算得到的相似搭配列表中的前 15 个搭配作为相似搭配集进行后续计算时，测试结果的准确率最高，这也能与前文计算的最相似 n 个搭配中改正搭配的覆盖率保持一致，取较多的相似搭配能覆盖改正搭配的几率更大。当取计算得到的相似搭配列表中的前 5 个搭配作为相似搭配集进行后续计算时，能够取得最高的召回率，这说明取较少的相似搭配对正确测试句更有利，由于减少了引入混淆的数目，因此有利于减少将正确句误改错的几率。根据 MRR 值可知，本模型基本能在最终建议结果列表的前两条中包含正确改正，取相似搭配计算的前 15 个搭配得到的 MRR 值最大，本文认为这与其高准确率的原因类似，因为其改正搭配的覆盖率更大。

上述结果也能表明提出的模型的易扩展性和实用性，只要向搭配库增加足够多的动名词搭配并训练这些搭配的分类器，理论上该模型就可用于纠正英语学习者实际写作中的动名词搭配错误。

结束语　本文通过分析中国英语学习者常见动名词搭配错误的产生原因和改正方法，构建了一个动名词搭配库，并提出了一种度量搭配间的相似度的方法，通过相似度计算来得到粗略的相似搭配集，然后使用每个搭配的分类器来过滤掉不能将测试句划为正确类的搭配进而得到候选结果集，最后使用三元语言模型对候选结果集进行打分排序得到最终的建议改正列表，经过实验发现该方法能取得很好的效果。

在未来的工作中，本模型所构建的搭配库规模需要进一步增大，可以通过将搭配库按相似性分组来避免检测目标和搭配库中的全部搭配做计算；需要为搭配库中的每个搭配收集更多的句子来训练对应分类器以提高过滤的精度；此外，本文处理动名词搭配错误的方法可以为解决其他种类的搭配错误提供借鉴意义。

参 考 文 献

[1] Yang Hui-zhong. Corpus-Based Analysis of Chinese Learner English[M]. Shanghai: Shanghai Foreign Language Education Press, 2005(in Chinese)
杨惠中. 基于 CLEC 语料库的中国学习者英语分析[M]. 上海：上海外语教育出版社，2005

[2] Liu L E, Wible D, Tsao N L. Automated Suggestions for Miscollocations[C]// Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications. Boulder, Colorado, June 2009: 47-50

[3] Wu Jian-cheng, Chang Y C, Mitamura T, et al. Automatic Collocation Suggestion in Academic Writing[C]// Proceeding of the ACL 2010 Conference Short Papers. Uppsala, Sweden, July 2010: 115-119

[4] Wang Rui. Theoretical Reflection and Reconstruction-A Corpus-based Study on Verb-noun Collocation Errors by Chinese non-English Majors from the Perspective of Conceptual Transfer [M]. Foreign Language Research, 2014(in Chinese)
王瑞. 理论反思与构拟——概念迁移视域中学习者动名词搭配错误序列分析之一 [M]. 外语学刊，2014

[5] Cao Li. Anslysis of Verb-noun Collocation Errors in College Students' CET-4 and CET-6 Examinations based on CLEC[D]. Wuhan: Huazhong University of Science and Technology, 2007 (in Chinese)
曹莉. 基于语料库的中国大学生英语四、六级考试作文中动名搭配错误分析[D]. 武汉：华中科技大学，2007

[6] Jiang J J, Conrath D W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy[C]// Proceeding of International Conference Research on Computational Linguistics (ROCLING X). Taiwan, 1997

[7] Deng Yao-chen, Xiao De-fa. A study of Chinese College Students' English Delexical Verb Collocation Model[M]. Foreign Languages and Their Teaching, 2005(in Chinese)
邓耀臣，肖德法. 中国大学生英语虚化动词搭配型式研究[M]. 外语与外语教学，2005

[8] Rozovskaya A, Roth D. Algorithm Selection and Model Adaptation for ESL Correction Tasks[C]// Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon, June 2011: 924-933

[9] Crammer K, Dekel O, Keshet J, et al. Online Passive-Aggressive Algorithms[J]. Journal of Machine Learning Research, 2006 (7): 551-585

(下转第 250 页)

数目的方法；然后对数据进行初步的聚类分析；最后，一个相似性度量测度被提出并被应用于组织多个较小的簇去代表每个实际的簇。实验分析展示本文提出的 McIB 算法能够有效解决“均匀效应”的影响，从而有效地挖掘非平衡数据集中的聚类模式；同时相比于其他聚类算法，McIB 算法在非平衡数据集上的数据分析性能在整体上表现更优。

参 考 文 献

[1] He H, Garcia E A. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284

[2] Longadge R, Dongre S. Class Imbalance Problem in Data Mining: Review[J]. International Journal of Computer Science and Network, 2013, 2(1): 83-87

[3] Chawla Nitesh V. Data mining for imbalanced datasets: An overview[M]// Data Mining and Knowledge Discovery Handbook, 2005. US: Springer, 2005: 853-867

[4] Provost F. Machine learning from imbalanced data sets 101 [C]// Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets, 2000. 2000: 1-3

[5] Zhi Wei-mei, Guo Hua-ping, Fan Ming, et al. Disscussion on classfication for imbalanced Data sets[J]. Computer Science, 2012, 39(S1): 304-308(in Chinese)
职为梅，郭华平，范明，等. 非平衡数据集分类方法探讨[J]. 计算机科学，2012，39(S1)：304-308

[6] Kumar C N S, Rao K N, Govardhan A, et al. Imbalanced K-Means: An algorithm to cluster imbalanced-distributed data[J]. International Journal of Engineering and Technical Research, 2014, 2(2): 114-122

[7] Jain A K, Dubes R C. Algorithms for clustering data[M]. Englewood Cliffs: Prentice hall, 1988

[8] Nguyen C H, Ho T B. An imbalanced data rule learner [C]// Knowledge Discovery in Databases: PKDD 2005. Berlin: Springer, 2005: 617-624

[9] MacQueen J. Some methods for classification and analysis of multivariate observations [C]// Proceedings of Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967. Oakland: CA, 1967: 281-297

[10] Abolkarlou N A, Niknafs A A, Ebrahimpour M K. Ensemble imbalance classification: Using data preprocessing, clustering algorithm and genetic algorithm[C]// 2014 4th International Conference on Computer and Knowledge Engineering (ICCKE). IEEE, 2014: 171-176

[11] Yen S-J, Lee Y-S. Cluster-based under-sampling approaches for imbalanced data distributions[J]. Expert Systems with Applications, 2009, 36(3): 5718-5727

[12] Zhang Y P, Zhang L N, Wang Y C. Cluster-based majority under-sampling approaches for class imbalance learning[C]// 2010 2nd IEEE International Conference on Information and Financial Engineering (ICIFE). IEEE, 2010: 400-404

[13] Xiong H, Wu J, Chen J. K-means clustering versus validation measures: a data-distribution perspective[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(2): 318-331

[14] Liang J, Bai L, Dang C, et al. The-Means-Type Algorithms Versus Imbalanced Data Distributions[J]. IEEE Transactions on Fuzzy Systems, 2012, 20(4): 728-745

[15] Qian J, Saligrama V. Spectral clustering with imbalanced data [C]// 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014: 3057-3061

[16] Prachuabsupakij W, Soonthornphisaj N. Cluster-based sampling of multiclass imbalanced data [J]. Intelligent Data Analysis, 2014, 18(6): 1109-1135

[17] Tishby N, Pereira F C, Bialek W. The information bottleneck method[C]// Proceedings of 37th Allerton Conference on Communication, Control and Computering, 1999. 1999: 368-377

[18] Cover T M, Thomas J A. Elements of information theory[M]. New York: John Wiley & Sons, 2012

[19] Slonim N. The information bottleneck: Theory and applications [D]. Jerusalem: The Hebrew University of Jerusalem, 2002

[20] Slonim N, Tishby N. Document clustering using word clusters via the information bottleneck method[C]// Proceedings of Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2000. ACM, 2000: 208-215

[21] Lou Zheng-zheng, Yang Chen, Ye Yang-dong. An IB algorithm based on data selection model[J]. Acta Electronica Sinica, 2014, 42(9): 1839-1846(in Chinese)
娄铮铮，杨晨，叶阳东. 基于数据选择模型的 IB 算法[J]. 电子学报，2014，42(9)：1839-1846

[22] DeGroot M H, Schervish M J, Fang X, et al. Probability and statistics[M]. MA: Addison-Wesley Reading, 1986

[23] Alcala-Fdez J, Sanchez L, Garcia S, et al. KEEL: a software tool to assess evolutionary algorithms for data mining problems[J]. Soft Computing, 2009, 13(3): 307-318

[24] Shi J, Malik J. Normalized cuts and image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905

[25] Yang Y. An evaluation of statistical approaches to text categorization[J]. Information Retrieval, 1999, 1(1/2): 69-90

[26] Manning C D, Raghavan P, Schütze H. Introduction to information retrieval[M]. Cambridge: Cambridge University Press, 2008

(上接第 233 页)

[10] Mariano F, Zheng Yuan. Grammatical error correction using hybrid systems and type filtering[C]// Proceedings of the 18th Conference on Computational Natural Language Learning. Baltimore, July 2014: 15-24

[11] Collins M. Modeling L. Course notes for NLP by Michael Collins [D]. Columbia University, Spring 2013

[12] Pauls A, Klein D. Faster and Smaller N-Gram Language Models [C]// Proceeding of 49th Annual Meeting of the ACL. Portland, Oregon, June 2011: 258-267

[13] Kneser R, Ney H. Improved backing-off for M-gram language modeling[C]// Proc. ICASSP. 1995: 181-184

[14] Craswell N. Mean Reciprocal Rank[M]. Springer US: Encyclopedia of Database Systems, 2009: 1703