基于差异性评估对 Co-training 文本分类算法的改进

唐焕玲[1,2],林正奎[1],鲁明羽[1]
(1. 大连海事大学信息科学技术学院,辽宁大连 116026; 2. 烟台职业学院计算机与信息工程系,山东烟台 264670)

摘 要: Co-training 算法要求两个特征视图满足一致性和独立性假设,但是,许多实际应用中不存自然的划分且满足这种假设的两个视图,且直接评估两个视图的独立性有一定的难度. 分析 Co-training 的理论假设,本文把寻找两个满足一致性和独立性特征视图的目标,转变成寻找两个既满足一定的正确性,又存在较大的差异性的两个基分类器的问题. 首先利用特征评估函数建立多个特征视图,每个特征视图包含足够的信息训练生成一个基分类器,然后通过评估基分类器之间的差异性间接评估二者的独立性,选择两个满足一定的正确性和差异性比较大的基分类器协同训练. 根据每个视图上采用的分类算法是否相同,提出了两种改进算法 TV-SC 和 TV-DC. 实验表明改进的 TV-SC 和 TV-DC 算法明显优于基于随机分割特征视图的 Co-Rnd 算法,而且 TV-DC 算法的分类效果要优于 TV-SC 算法.

关键词: 半监督文本分类; Co-training; 特征视图; 差异性评估; 标注文本; 未标注文本
中图分类号: TP181 **文献标识码**: A **文章编号**: 0372-2112 (2008) 12A-138-06

An Improved Co-training Text Categorization Algorithm Based on Diversity Measures

TANG Huan-ling[1,2], LIN Zheng-kui[1], LU Ming-yu[1]
(1. *College of Information and Science Technique, Dalian Maritime University, Dalian, Liaoning* 116026, *China*;
2. *Department of Computer and Information Engineering, Yantai Vocational College, Yantai, Shandong* 264670, *China*)

Abstract: Co-training algorithm is constrained by its assumption that the features can be split into two compatible and independent subsets. However, the assumption is usually violated in real-world application, especially for independence. We discover its real purpose is to find two classifiers with certain accuracy and sufficient diversity to co-train. First, multi-views are created using different term evaluation functions. Second, instead of directly computing the independence between two sub-views, this paper evaluates the independence between two classifiers, trained on them, by using diversity measures indirectly. Thus a pair of classifiers with certain accuracy and greater diversity is selected. The experimental results show two improved algorithms named TV-SC and TV-DC are both outperform another co-training algorithm named Co-Rnd based on random splitting method, and TV-DC outperforms TV-SC.

Key words: semi-supervised text categorization; Co-training; features views; diversity measures; labeled documents; unlabeled documents

1 引言

传统文本分类算法需要大量标注样本,但是已标注的样本数量有限,获取高质量的标注样本代价昂贵. 于是利用少量的标注样本和大量的未标注样本的半监督学习(Semi-Supervised Learning)引起了广泛关注[1~7]. Co-training 算法[2,3]是一种经典的半监督学习算法,假设数据集可以被自然地分成两个独立的特征视图,每个视图都包含足够的信息进行分类学习,在每个视图上建立各自的分类器,两个分类器每次互相标记一部分置信度高的样本给对方,重新训练,直到没有更多适合的未标记样本加入.

Co-training 的理论假设是:数据集可以被自然地分成两个独立的特征视图,并要求两个特征视图满足一致性(compatible)和独立性(independent). 前者表示,对大多数样本目标函数在每个特征子集上预测的类别是完全相同的. 后者表示对指定类别的任意的样本,在两个视图中的描述是独立的[2,3].

由于多种原因,这两个假设并不能完全严格地满足,尤其是独立性,甚至在很多情况下不存在自然分割的两个视图,如文本分类的训练文本集,这就制约了 Co-training 算法的有效运用. Nigam 与 Ghani 人工随机将

收稿日期:2008-08-07;修回日期:2008-11-28
基金项目:国家自然科学基金(No. 60773084,J0724003,60603023);教育部博士点基金(No. 20070151009)

特征集合分割成两个子集[3];Zhou 与 Goldman 提出了在单个特征视图上建立多个分类器的 Democratic Co-learning 算法[5];Zhou 和 Li 提出了使用三个分类器的 Tri-training 算法[6]. 这些方法在一定程度上放松了对 Co-training 的假设约束,没有解决 Co-training 的理论假设问题.

分析 Co-training 的理论假设,要求两个特征视图满足一致性和独立性,最终目的是为了生成两个满足一定的正确性(accuracy),且存在较大差异性(diversity)两个基分类器进行协调训练. 因为合理的差异,能够减少两个基分类器给同一个未标注样本都标注错误的可能性. 直接评估两个特征视图的独立性有一定的难度,但评估两个基分类器的正确性和差异性比较容易. 因此,我们把寻找两个满足一致性的和独立性的特征视图的目标,转变为寻找两个既满足一定的正确性,又存在较大的差异性的两个基分类器的问题.

鉴于此,本文提出了一种基于差异性评估对 Co-training 算法的改进方案. 首先基于我们的 TEF-WA (Term Evaluation Function-Weight Adjustment)技术[8],用特征评估函数建立多个特征视图,生成多个基分类器. 然后在基分类器满足一定的正确性前提下,用多种方法评估两个分类器之间的差异性,实现间接地评估二者之间的相关性. 最后,选择两个既满足一定的正确性,又存在较大的差异性的基分类器协同训练,提高 Co-training 分类的效果.

2　基于 TEF-WA 技术建立多视图

文本的表示采用向量空间法(VSM,Vector Space Model),文本 x_i 用特征向量$(w_{i1}, w_{i2}, \cdots, w_{ik}, \cdots, w_{im})$表示,$w_{ik}$表示 x_i 的第 k 个特征 t_{ik}的权重. 特征选择和权重调整是文本分类的关键步骤,直接影响分类器的分类结果. 权重计算常用的是 TF-IDF 公式[11,12],我们采用 TEF-WA 技术[8],即利用信息论中常用的评估函数代替逆文本频率(IDF)给每个特征独立的打分,评估分的高低能够很好地代表特征的重要性,根据评估分调整特征的权重,如式(1).

$$w_{ik} = TF - TEF(t_{ik}) = TF(t_{ik}) * TEF(t_{ik}) \quad (1)$$

$TF(t_{ik})$表示文本 x_i 的第 k 个特征的词频,$TEF(t_{ik})$表示常用的评估函数,用于给各个特征词打分,反映特征词与各类之间的相关程度. 如文本频率(Document Frequency,DF),信息增益(Information Gain,IG),期望交叉熵(Expected cross Entropy,ECE),互信息(Mutual Information,MI),文本证据权(Weight of Evidence for Text,WET),几率比(Odds Ratio),χ^2 统计量(CHI)等[8,11,12]. 这里使用评估函数不仅是为了调整权重,更重要的是利用不同的评估函数创建不同的特征视图,训练生成有差异的基分类器.

(1) 文本频率

$freq(t, c)$表示特征词 t 在 c 类文本中出现的频率.

$$TEF_{df}(t) = freq(t, c) \quad (2)$$

(2) 信息增益

信息增益衡量的是特征词 t 在一个文本中出现或不出现时所获得的信息的比特数.

$$TEF_{InfoGain}(t) = P(t)\sum_j P(c_j|t)\log\frac{P(c_j|t)}{P(c_j)} + P(\bar{t})\sum_j P(c_j|\bar{t})\log\frac{P(c_j|\bar{t})}{P(c_j)} \quad (3)$$

t 表示特征词出现,$\bar{t}$ 表示特征词 t 不出现. $P(t)$表示特征词 t 出现的概率,$P(\bar{t})$表示特征词 t 不出现的概率. $P(c_j)$是类 c_j 的先验概率,$P(c_j|t)$是基于 t 的类 c_j 的条件概率.

(3) 期望交叉熵

期望交叉熵衡量的是特征词 t 在特征文本中出现时所获得的信息量,与信息增益不同.

$$TEF_{CrossEntry}(t) = P(t)\sum_j P(c_j|t)\log\frac{P(c_j|t)}{P(c_j)} \quad (4)$$

(4) 互信息

互信息衡量某个特征词和类别之间的统计独立关系,对特征词 t 和某个类别 c_j 的互信息定义如下:

$$TEF_{MutualInfo}(t) = \sum_j P(c_j)\log\frac{P(t|c_j)}{P(t)} \quad (5)$$

(5) 文本证据权

文本证据权衡量的是类的概率和给定特征时类的条件概率之间的差别,不需要计算 t 的所有可能值,而只考虑 t 在文本中是否出现.

$$TEF_{WeightEvidTxt}(t) = P(t)\sum_j P(c_j)\left|\log\frac{P(c_j|t)(1-P(c_j))}{P(c_j)(1-P(c_j|t))}\right| \quad (6)$$

(6) 几率比

$$TEF_{OddsRatio}(t) = \log\frac{P(t|pos)(1-P(t|pos))}{P(t|neg)(1-P(t|neg))} \quad (7)$$

pos 代表正类,neg 代表负类,$P(t|pos)$表示特征 t 在正类 pos 中出现的概率,$P(t|neg)$表示 t 在负类 neg 中出现的概率,特别适用于二元分类器.

(7) χ^2 统计量(CHI)

χ^2 统计量用于衡量一个特征词和一个类别之间的统计独立关系. 令 a 为训练集中包含 t 的 c_j 类文本数,b 为包含 t 的非 c_j 类文本数,d 为不包含 t 的 c_j 类文本数,e 为不包含 t 的非 c_j 类文本数,N 为总文本数. 特征词 t 和类别 c_j 之间的 χ^2 统计量定义为:

$$TEF_{\chi^2}(t) = \sum_j P(c_j)\frac{N(ae-bd)^2}{(a+d)(b+e)(a+b)(d+e)} \quad (8)$$

根据上述评估函数,在同一个训练集上构造不同的特征视图. 令 L 表示带类别标注的训练文本集,V_t 表示根据某个评估函数建立的特征视图.

$$V_t = splt\_view(TEF, TForDF, m) \quad (9)$$

TEF 表示特征评估函数,TForDF 表示特征权重计算使用的是词频型(TF)还是文档型(DF)公式[8]. 文档频数型公式不考虑一个特征词在一个文本中出现的次数,只考虑它是否出现过. 词频型公式则考虑特征词在文本中出现的次数. m 表示保留特征数量.

例如,我们用 MI/DF/1000 表示选择使用评估函数 MI、文本型公式、保留 1000 特征词建立特征视图,而 Odds/TF/900 代表使用评估函数 Odds、词频型公式、保留 900 特征词建立的另一个特征视图. 对同一个文本 x_i 来说,其特征向量在不同的视图上是不同的,即便是同一种分类算法在不同的特征视图上训练生成的分类器也会有差异.

3 两个基分类器之间的差异性评估

通过分析 Co-training 的理论假设,把寻找两个满足较高一致性和独立性特征视图目标,转变为寻找两个满足一定的正确性和较大的差异性的基分类器的问题. 评估基分类器的正确性比较简单,本文采用宏平均精度 Macro-P,如式(10)所示. $TP(c_j)$ 表示属于 c_j 类,且被正确分为 c_j 类的样本数; $FP(c_j)$ 为不属于 c_j 类,但是被分为 c_j 类的样本数.

$$Macro\_P = \frac{1}{|C|}\sum_{j=1}^{|C|}\frac{TP(c_j)}{TP(c_j) + FP(c_j)} \quad (10)$$

这里重点讨论基分类器间的差异性评估. 集成分类学习通常使用 PMD(Pairwise Diversity Measures)[9,10] 评估每对基分类器之间的差异性,这里利用属于 PMD 的 Q 统计、相关系数 ρ、不一致性和双误法[9,10] 评估 Co-training 的两个基分类器之间的差异性.

令 $H = \{h_1, \cdots, h_M\}$ 表示一组基分类器,$D = \{(x_1, y_1), \cdots, (x_N, y_N)\}$ 表示一组带类标签的样本. 基分类器 h_t 的分类输出对应一个 N 维向量 $\mu_t = [\mu_{t1}, \cdots, \mu_{ti}, \cdots, \mu_{tN}]^{\mathrm{T}}$, $t = 1, \cdots, M$. μ_{ti} 表示 h_t 对样本 x_i 的分类结果,分对为 1,分错为 0. 令 N^{11} 表示 h_t 和 h_s 都分类正确的样本数,N^{01} 表示 h_t 分类错误而 h_s 分类正确的样本数,N^{10} 表示 h_t 分类正确而 h_s 分类错误的样本数,N^{00} 表示 h_t 和 h_s 都分类错误的样本数,则:

$$N^{11} = \sum_{i=1}^{N}(\mu_{ti} \wedge \mu_{si}), \quad N^{01} = \sum_{i=1}^{N}(\bar{\mu}_{ti} \wedge \mu_{si})$$

$$N^{10} = \sum_{i=1}^{N}(\mu_{ti} \wedge \bar{\mu}_{si}), \quad N^{00} = \sum_{i=1}^{N}(\bar{\mu}_{ti} \wedge \bar{\mu}_{si})$$

①Q 统计:对基分类器 h_t 和 h_s,Q 统计定义如式(11):

$$Q_{t,s} = (N^{11}N^{00} - N^{01}N^{10}) / (N^{11}N^{00} + N^{01}N^{10}) \quad (11)$$

如 h_t 和 h_s 统计独立,那么 $Q_{t,s} = 0$, $Q_{t,s} \in [-1, 1]$. 两个基分类器的分类趋向一致 $Q_{t,s} > 0$,否则 $Q_{t,s} < 0$.

②相关系数 ρ:评估 h_t 和 h_s 之间的相关性用 $\rho_{t,s}$ 度量,如式(12):

$$\rho_{t,s} = \frac{(N^{11}N^{00} - N^{01}N^{10})}{\sqrt{(NN^{11} + N^{10})(N^{01} + N^{00})(N^{11} + N^{01})(N^{10} + N^{00})}} \quad (12)$$

如果 $\rho_{t,s} = 0$,那么两个基分类器 h_t 和 h_s 完全不相关,$\rho_{t,s}$ 越大二者的相关度越大. $\rho_{t,s} \in [0, 1]$.

③不一致法:$Dis_{t,s}$ 表示 h_t 和 h_s 的不一致如式(13):

$$Dis_{t,s} = (N^{01} + N^{10}) / (N^{11} + N^{01} + N^{10} + N^{00}) \quad (13)$$

④双误法:$DF_{t,s}$ 表示 h_t 和 h_s 都分错的概率,如式(14):

$$DF_{t,s} = N^{00} / (N^{11} + N^{01} + N^{10} + N^{00}) \quad (13)$$

⑤综合法:用 $DM(h_t, h_s)$ 表示综合评估法如式(15):

$$DM(h_t, h_s) = \alpha Q_{t,s} + \beta\rho_{t,s} + \gamma(1 - Dis_{t,s}) + \delta DF_{t,s} \quad (15)$$

其中,$\alpha + \beta + \gamma + \delta = 1$, $\alpha, \beta, \gamma, \delta \in [0, 1]$.

4 改进的算法 TV-SC 与 TV-DV

根据每个视图上采用的分类算法是否相同,分别记作 TV-SC(Two Views-Same Classifiers)和 TV-DC(Two Views-Different Classifiers),如表 1 和表 2 所示. L 表示标注样本集,U 表示未标注样本集,$C = \{c_1, \cdots, c_K\}$ 表示一组类标记,$V = \{V_1, \cdots, V_M\}$ 表示多个特征视图,$H = \{h_1, \cdots, h_M\}$ 表示在 V 上建立的多个基分类器. f_1, f_2 表示两种不同的分类算法,可以是 Naive Bayes(NB)、K 近邻(KNN)、质心向量法(CenVSM)、支持向量机(SVM)等. $DM(h_t, h_s)$ 表示基分类器 h_t 和 h_s 之间的差异性. $Mp(h_t)$ 和 $Mp(h_s)$ 表示 h_t 和 h_s 的宏平均分类精度,r 是迭代次数.

表 1 TV-SC 算法(Two Views-Same Classifiers algorithm)

1) Create M feature views $V_1, \cdots, V_M$ based TEF-WA; (see Equ. 1～9);
2) Use f and $V_t(L)$ to create classifiers h_t, $t = 1, \cdots, M$;
3) Compute $Mp(h_t)$, $Mp(h_s)$ and $DM(h_t, h_s)$, $t, s = 1, \cdots, M$; (See Equ. 10～15);
4) Select two classifiers with certain accuracy and higher diversity according to $Mp(h_t)$, $Mp(h_s)$ and $\{DM(h_t, h_s)\}$, let V_1 and V_2 be the associated sub-views;
5) Loop for r iterations
  5.1) Create classifiers h_1 and h_2 using f and $V_1(L)$, $V_2(L)$ respectively;
  5.2) For each class c_j Do
    5.2.1) Let $b1$ and $b2$ be unlabeled documents on which h_1 and h_2 make

most confident prediction for c;

5.2.2) Remove $b1$ and $b2$ from U, label them according to h_1 and h_2, and add them to L respectively;

6) Combine the prediction of h_1 and h_2.

表 2　TV-DC 算法(Two Views-Different Classifiers algorithm)

1) Create M feature views V_1, …, V_M based TEF-WA; (see Equ. 1～9);

2) Use f_1 and $V_t(L)$ to create classifiers h_{t1}, Use f_2 and $V_s(L)$ to create classifiers h_{s2}, $t, s = 1, \cdots, M$;

3) Compute $Mp(h_{t1})$, $Mp(h_{s2})$ and $DM(h_{t1}, h_{s2})$, $t, s = 1, \cdots, M$; (See Equ. 10～15);

4) Select two classifiers with certain accuracy and higher diversity according to $Mp(h_{t1})$, $Mp(h_{s2})$ and $\{DM(h_{t1}, h_{s2})\}$, let V_i and V_j be the associated sub-views;

5) Loop for r iterations

5.1) Create classifiers h_1 and h_2 using f_1 and $V_i(L)$, f_2 and $V_j(L)$ respectively;

5.2) For each class c Do

5.2.1) Let $b1$ and $b2$ be unlabeled documents on which h_1 and h_2 make most confident prediction for c;

5.2.2) Remove b_1 and b_2 from U, label them according to h_1 and h_2, and add them to L respectively;

6) Combine the prediction of h_1 and h_2.

5　实验分析

实验数据采用从易宝中文下载的中文新闻文本作数据集,包含 20341 篇分属于经济、政治、国际、文教和体育五大类别的新闻,整个数据集被划分成多个不同的子集. 分类结果采用宏平均精度 Macro-P、宏平均召回率 Macro-R、宏平均 F1 值、Macro-F1 和微平均 F1 值 Micro-F1 评估.

为了验证 TV-SC 和 TV-DC 的分类效果,与基于随机分割特征视图的 Co-training 算法(记作 Co-Rnd)进行了实验比较. 表 3 和表 4 所示是 200 篇标注样本和 500 篇未标注样本上 TV-DC、TV-SC 以及 Co-Rnd 算法的分类结果,计算 DM 时,$\alpha = \beta = \delta = 0.333$, $\gamma = 0$.

从表 3 可以看出: ①TV-DC 的分类效果不仅与每个基分类器的正确性有关,还与基分类器间的差异性有关. 第 1 组数据表明由参数 ECE/ DF/ 1300(NB)和 Odds/ TF/ 900(CenVSM)构建的一对基分类器间的差异性最大, Q、ρ、DF 和 DM 值最小. 虽然 Odds/ TF/ 900(CenVSM)生成的基分类器的精度仅 73.52 %,但是由于两个基分类器的差异性最大,使得 TV-DC 的分类效果仅次于第 2 组,Macro-P 达到 83.86 %. 第 2 组的一对基分类器的正确性相对其它 3 组最好,差异性指标 DM 次于第 1 组,同时考虑基分类器的正确性和差异性,第 2 组的一对基分类器最好,二者构成的 TV-DC 分类器的分类效果也最好,Macro-P 达到了 85.25 %. 比较第 3、第 4 组数据,说明当两个基分类器的分类正确性差不多时,二者的差异性越大, TV-DC 的分类效果越好. ②Q、ρ、Dis、DF 方法都能反映基分类器间的差异性,但综合指标 DM 最好. ③比较每组数据 TV-DC 分类器是否使用未标注文本的上下两行,使用未标注文本能够明显提高分类效果,Macro-P 提高了 4.46 %～5.97 %, Micro-F1 提高了 3.48 %～5.22 %, 表明 TV-DC 算法结合未标注文本能明显地提高分类效果.

表 3　TV-DC 算法分类效果以及两个基分类器的分类正确性和差异性比较

No.	Sub-View (Classifier)	Macro-P (%)	Diversity Measures					L + U	Results of TV-DC Classifier (%)			
			Q	*ρ*	*Dis*	*DF*	*DM*		Macro-P	Macro-R	Macro-F1	Micro-F1
1	ECE/ DF/ 1300 (NB) Odds/ TF/ 900 (CenVSM)	**81.09** 73.52	**0.2874**	**0.1013**	**0.5826**	**0.1739**	**0.19**	yes No	**83.86** 77.96	**82.89** 77.80	**82.89** 77.30	**83.48** 78.26
2	MI/ TF/ 1000 (NB) Odds/ TF/ 1200 (CenVSM)	**79.77** **78.19**	**0.3399**	**0.1305**	0.5391	**0.1913**	**0.22**	yes No	**85.25** 80.79	**83.44** 80.34	**83.50** 80.62	**84.35** 80.87
3	IDF/ DF/ 900 (NB) Odds/ TF/ 900 (CenVSM)	78.92 73.52	0.3450	0.1246	0.5652	0.1913	0.22	yes No	82.96 76.98	81.89 76.96	81.80 76.50	82.61 77.39
4	IG/ TF/ 1200 (NB) Odds/ TF/ 1000 (CenVSM)	77.77 73.52	0.4773	0.1700	0.5478	0.2000	0.28	yes No	81.36 75.95	80.15 76.06	80.26 75.74	80.87 76.52

表 4 TV-DC、TV-SC 与 Co-Rnd 的分类结果以及基分类器之间的正确性和差异性比较

No.	Classifier	Sub-View(Classifier)	Macro-P (%)	Diversity Measures					Macro-P (%)	Macro-R (%)	Macro-F1 (%)	Micro-F1 (%)
				Q	*ρ*	*Dis*	*DF*	*DM*				
1	TV-DC	MI/ TF/ 1000(NB) Odds/ TF/ 1200(CenVSM)	79.77 78.19	**0.3399**	**0.1305**	0.5391	0.1913	**0.22**	**85.25**	**83.44**	**83.50**	**84.35**
	TV-SC	MI/ TF/ 1000(NB) Odds/ TF/ 1200(NB)	79.77 80.56	0.4531	0.1582	0.5478	0.1652	0.26	78.71	77.06	77.21	78.26
	Co-Rnd	RNDV1(NB) RNDV2(CenVSM)	75.47 73.75	0.5464	0.2788	0.3304	0.1826	0.34	74.07	72.38	70.87	72.38
2	TV-DC	ECE/ DF/ 1300(NB) Odds/ TF/ 900(CenVSM)	81.09 73.52	**0.2874**	**0.1013**	0.5826	0.1739	**0.19**	**83.86**	**82.89**	**82.89**	**83.48**
	TV-SC	ECE/ DF/ 1300(NB) Odds/ TF/ 900(NB)	**81.09** **80.25**	0.3419	0.1130	0.5913	0.1565	**0.20**	**82.23**	80.93	81.09	81.74
	Co-Rnd	RNDV1(NB) RNDV2(CenVSM)	78.26 75.74	0.5392	0.2757	0.3304	0.1826	0.33	77.78	75.00	76.36	76.51
3	TV-DC	IDF/ DF/ 900(NB) Odds/ TF/ 900(CenVSM)	78.92 73.52	0.3450	0.1246	0.5652	0.1913	**0.22**	**82.96**	81.89	81.80	82.61
	TV-SC	IDF/ DF/ 900(NB) Odds/ TF/ 900(NB)	78.92 80.25	0.4010	0.1366	0.5739	0.1739	0.24	80.33	79.09	79.16	80.00
	Co-Rnd	RNDV1(NB) RNDV2(CenVSM)	72.70 71.43	0.5172	0.2658	0.3391	0.1913	0.32	73.52	73.07	72.67	73.91

表 4 是 TV-DC、TV-SC 和 Co-Rnd 算法的分类比较. TV-DC 在每个视图上的采用不同分类算法(NB 和 CenVSM)建立基分类器;而 TV-SC 在两个视图上采用相同分类算法 NB. Co-Rnd 算法基于随机分割法生成两个特征视图,且两个视图上分别采用 NB 和 CenVSM 分类算法.

从表 4 可以看出,TV-DC 的分类效果优于 TV-SC,Macro-P 提高 1.63 %～6.54 %,Macro-R 提高 1.96 %～6.38 %. 因为不同分类算法在同一对特征视图上,建立的两个基分类器间的差异性,比相同分类算法建立的基分类器的差异性大.

第 2 组数据 TV-SC 算法下两个基分类器的 Q、ρ、DF 和 DM 的值都比较小,即二者差异性比较大. 说明调整式(9)的参数,能找到一对独立性较强的特征视图,且二者分类精度也比较高,因此 TV-SC 的分类效果也比较好,Macro-P 为 82.23 %.

Co-Rnd 采用随机分割视图法建立的两个基分类器间的差异性较小,分类结果明显比 TV-SC 和 TV-DC 差. 对比 Co-Rnd,TV-DC 的 Macro-P 提高 6.08 %～10.18 %,TV-SC 的 Macro-P 提高 4.45 %～6.81 %,Macro-R、Macro-F1 和 Micro-F1 的值也表明 TV-DC 和 TV-SC 的分类结果明显优于 Co-Rnd.

6 结论

通过分析 Co-training 的理论假设,把寻找两个满足较高一致性和独立性特征视图的目标,转变为寻找两个满足一定的正确性和较大的差异性的基分类器的问题,提出了基于差异性评估对 Co-training 的改进算法 TV-SC 和 TV-DC. 首先基于 TEF-WA 技术利用特征评估函数及其它参数建立多个特征视图,然后利用多种方法计算基分类器之间差异,最后选择两个既满足一定的正确性,又存在较大的差异性的基分类器协同训练,减少给同一个未标注文本都标注错误的可能性,从而提高 Co-training 分类的效果.

参考文献:

[1] Seeger M. Learning with labeled and unlabeled data[R]. University of Edinburgh, Edinburgh, UK 2001.

[2] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[A]. In Proceedings of the Workshop on Computational Learning Theory[C]. New York: ACM Press, 1998. 92 - 100.

[3] Nigam K, Ghani R. Analyzing the effectiveness and applicability of co-training[A]. In Proceedings of ninth International Conference on Information and Knowledge Management[C]. New York: ACM Press, 2000. 86 - 93.

[4] Balcan M-F, Blum A. A PAC-style model for learning from labeled and unlabeled data[A]. In Proceedings of the 18th Annual Conference on Learning Theory[C]. Berlin Heidelberg: Springer-Verlag, 2005. 111 - 126.

[5] Zhou Y, Goldman S. Democratic co-learning[A]. In Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence[C]. Washington, DC: IEEE Computer Society Press, 2004. 594 - 602.

[6] Zhou Z-H, Li M. Tri-training: exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data

Engineering,2005,17(11):1529 - 1541.

[7] Chapelle O,Sindhwani V,Keerthi S S. Optimization techniques for semi-supervised support vector machines[J]. Journal of Machine Learning Research,2008,9:203 - 233.

[8] 唐焕玲,孙建涛,陆玉昌. 文本分类中结合评估函数的 TEF-WA 权值调整[J]. 计算机研究与发展,2005,42(1):47 - 53.
Tang Huanling,Sun Jiantao,Lu Yuchang. A weight adjustment technique with feature weight function named TEF-WA in text categorization[J]. Journal of Computer Research and Development,2005,42(1):47 - 53. (In Chinese)

[9] Kuncheva L I,Whitaker C J. Measures of diversity in classifier ensembles[J]. Machine Learning,2003,51(2):181 - 207.

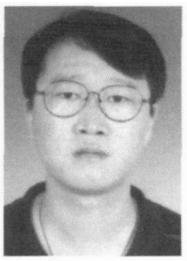
[10] Ruta D,Gabrys B. A theoretical analysis of the limits of majority voting in multiple classifier systems[J]. Pattern Analysis & Applications,2002,5(4):333 - 350.

[11] Yang Y,Pedersen J P. A comparative study on feature selection in text categorization[A]. In Proceedings of the Fourteenth International Conference on Machine Learning[C]. San Francisco,USA:Morgan Kaufmann Publishers,1997. 412 - 420.

[12] Fabrizio Sebastiani. Machine learning in automated text categorization[J]. ACM Computing Surveys,2002,34(1):1 - 47.

作者简介:

唐焕玲　女,副教授、博士生,1970 年生于山东龙口. 2004 年于清华大学获得工学硕士学位. 研究方向为文本挖掘和网络挖掘.
E-mail:thl01@163.com



林正奎　男,博士、副教授,1971 年生于山东烟台. 2005 年于大连理工大学获得工学博士学位. 研究方向为大型复杂信息系统方法及应用研究.

鲁明羽　男,教授、博士生导师,中国计算机学会高级会员. 1963 年生于黑龙江鸡西. 1988 年、2002 年于清华大学分别获得工学硕士和工学博士学位,研究方向为数据挖掘和机器学习.