

文章编号:1001-9081(2005)09-2028-03

基于 CBR 的文本自动分类研究

张婷慧,耿焕同,蔡庆生

(中国科学技术大学 计算机科学与技术系,安徽 合肥 230027)

(zth@mail.ustc.edu.cn)

摘 要:KNN 方法是性能最好的文本分类方法之一,但它在分类时要计算待分类文档与所有训练样本的相似度,时间复杂度较大。文中提出了一种基于 CBR 的文本自动分类方法,先用聚类方法把训练样本库转换为范例库,然后用 KNN 思想分类。实验结果显示该方法分类的平均召回率和准确率达到了 87.07% 和 89.17%;并且通过分析算法的时间复杂度得知,该方法的分类速度比 KNN 方法有很大的提高,因此具有很好的实用价值。

关键词:基于范例推理;文本自动分类;K 近邻;聚类

中图分类号:TP391 **文献标识码:**A

Study of automatic text categorization based on CBR

ZHANG Ting-hui, GENG Huan-tong, CAI Qing-sheng

(Department of Computer Science and Technology, University of Science and Technology of China, Anhui Hefei 230027, China)

Abstract: K-Nearest Neighbor (KNN) is one of the top-performing classifiers, but it has a large time complexity on calculating the similarity between the document and all training samples. An automatic text categorization mechanism based on CBR was presented, the training sample library was converted to the case library and the document was classified by KNN. In experiments, the average recall and precision were 87.07% and 89.17% respectively. In addition, by analyzing the time complexity, this mechanism can perform much more quickly than the KNN method.

Key words: case-based reasoning (CBR); automatic text categorization; K-nearest neighbor; clustering

0 引言

随着 Internet 的迅猛发展,在线文档信息的迅速增加,文本自动分类已经成为现代信息处理研究的一大热点。文本自动分类是在给定的分类体系下,根据文档的内容或属性,将大量文档归到一个或多个类别中的过程,召回率和准确率是文本分类两个常用的评价指标。常用的分类方法有支持向量机 (SVM)、K 近邻 (KNN)、神经网络 (Nnet)、线性最小二乘方估计 (LLSF) 和贝叶斯算法 (Bayes) 等^[1]。其中 KNN 方法实现简单、有很高的召回率和准确率,并且在待分类文档数从几百到几万时都有很好的分类效果^[1],因此,在文本自动分类领域有着广泛的应用。但 KNN 方法是一种基于要求的学习算法,它存放所有的训练样本,直到测试样本需要分类时才建立分类,当与测试样本比较的训练样本较多时,会导致很高的时间开销。

为了减小 KNN 方法的计算量,学者们进行了广泛的研究^[2-4]。这些方法的基本思想是在原来的训练样本中选取一些代表作为新的训练样本,或者删除原来训练样本中的某些样本,从而达到减小训练样本集的目的。但是这些方法在训练样本集中每增加或删除一个样本时,都有一个反复迭代至样本集不再变化的过程,在样本集很大时工作量非常大。并且若遇到无法分类的文档,需要对分类器进行调整时,计算量也很大。

基于范例推理(case-based reasoning, CBR)是一种机器学习

方法,它与人的思维方式类似,通过判断新问题与已解决问题的相似性,找到与之相似的问题的解,并对该问题的解进行适当地修正来解决新问题^[5]。本文在 CBR 框架下,借鉴 KNN 的思想,提出了一种基于 CBR 的文本自动分类方法,其基本思想为:首先将各类的训练样本聚类成簇,然后根据每个簇覆盖的训练样本计算出其特征向量,把一个簇作为类的一个范例。推理过程即分类过程,将待分类文档与所有范例比较,得到最近邻,根据最近邻计算其所属类别。由于范例的数量要远少于训练样本的数量,因此分类时计算量会减小,从而分类速度也会提高。根据以上方法实现了一个文本自动分类系统,并对其进行了测试,分类的平均召回率和准确率分别达到了 87.07% 和 89.17%。同时,分析算法的时间复杂度可知,与 KNN 方法相比,本方法大大减少了分类时间,提高了分类速度,因此具有很好的实用价值。

1 基于 CBR 的文本自动分类方法

基于 CBR 的文本自动分类方法的基本思想为:先运用聚类分析方法将各类的训练样本聚类成簇,然后根据每个簇覆盖的训练样本计算出其特征向量,把一个簇作为类的一个范例,这样将训练样本库转换成范例库。分类时,将待分类文档 X 与所有范例比较,得到与之最相似的 k 个范例,根据这 k 个范例决定 X 的类别。

1.1 范例表示

范例表示的过程即分类的训练过程。遍历训练样本库,

收稿日期:2005-04-01;修订日期:2005-07-22

基金项目:国家自然科学基金资助项目(70171052);皖泰开发项目资助(143-150401)

作者简介:张婷慧(1982-),女,硕士研究生,主要研究方向:人工智能、中文信息处理;耿焕同(1973-),男,博士研究生,主要研究方向:人工智能、知识发现;蔡庆生(1938-),男,教授,博士生导师,主要研究领域:人工智能、机器学习、知识发现。

按图1所示的方法分别对每个类别训练。

1.1.1 预处理

对每个训练样本,首先要进行预处理。根据标点符号将文档分成句子串,若其为中文文档,还需先进行分词,然后根据停用词表去掉停用词。在预处理之后,一篇文档可以被看作词的集合 $\{t_1, t_2, \dots, t_n\}$,其中 t_i 是词。同时,我们还记录每个句子对应的词的集合以及每个词所在的句子集合。

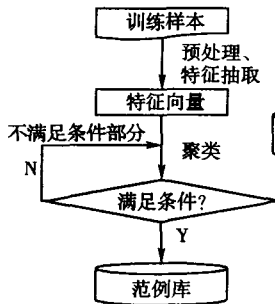


图1 每类的训练过程

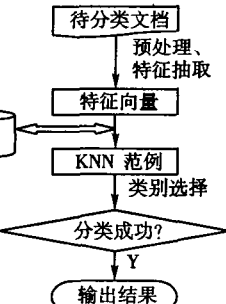


图2 分类过程

1.1.2 特征表示

对预处理后的文档,统计出每个词在文档中出现的频率,将该频率作为词的初始权重,然后用 TF-IDF 权重法^[6]来计算词的权重,取权重最大的 n 个词作为文档的特征项。TF-IDF 权重计算方法如公式1所示。

$$w_{ij} = \frac{[\log(f_{ij}) + 1.0] \times \log(N/n_j)}{\sqrt{\sum_{j=1}^m \{[\log(f_{ij}) + 1.0] \times \log(N/n_j)\}^2}} \quad (1)$$

其中 f_{ij} 是特征项 T_j 在文档 d_i 中出现的频率, N 为 d_i 所在类的文档总数, n_j 为该类中包含 T_j 的文档数, $\log(N/n_j)$ 即逆词频idf。

得到文档的特征项和权重以后,文档就可以用向量空间中的一个点来表示,形如 (W_1, W_2, \dots, W_n) ,其中 W_i 为第 i 个特征项的权重。向量 d_i, d_j 之间的相似度用余弦公式(公式2)来计算:

$$\text{sim}(d_i, d_j) = \frac{\sum_{k=1}^m w_{ik} \times w_{jk}}{\sqrt{(\sum_{k=1}^m w_{ik}^2) \cdot (\sum_{k=1}^m w_{jk}^2)}} \quad (2)$$

1.1.3 范例生成

为了减小分类时的计算量,我们将大量的训练样本抽象为较少的范例。每个范例代表一些训练样本,同时范例之间的关联要尽可能小。聚类分析方法^[7]是根据客体属性对一系列未分类客体进行类别的识别,目的是根据某种规则,将样本空间的样本数据集划分为表示不同模式或系统行为的一些子集(簇)。聚类方法使得簇间的相似性尽量少,而簇内的相似性尽量大。因此,我们用聚类分析方法把训练样本抽象成范例。

由于本文训练样本库比较大,选择 CLARA(Cluster Large Application)算法^[7]作为聚类算法。为了使每个范例中包含的训练样本数尽量平均,设置了一个阈值 maxnum,表示一个簇中训练样本的最大篇数,若簇中的训练样本数超过 maxnum,要对该簇再聚类,直到每个簇中的文档数都不超过 maxnum 为止。聚类得到的一个簇对应一个范例。遍历整个训练样本库,对每个类的训练样本进行聚类,就将训练样本库转换成了用于分类的范例库。范例库生成的过程如算法1所示:

算法1:范例库生成

输入:训练样本库 SampleBase, 阈值 maxnum

输出:范例库 CaseBase

step1: 取 SampleBase 中的一个类的文档,对其用 CLARA 算法聚类,聚类结果加入簇的集合 ClusterSet,并将该类的文档从 SampleBase 中删除;

step2: 若 ClusterSet 非空,取其中的一个簇 cluster;

step2.1: 如果 cluster 中的文档数 $|cluster| \leq \text{maxnum}$,根据 cluster 生成范例 case 加入范例库中。

设 cluster 中有 m 篇样本 $S_1, S_2, \dots, S_m, C': (W_1', W_2', \dots, W_n') = S_1 + S_2 + \dots + S_m$,生成范例 $C: (W_1, W_2, \dots, W_n)$

为对 C' 进行归一化的结果:

$$w_i = \frac{w_i'}{\sqrt{\sum_{j=1}^n w_j'^2}} \quad i = 1, \dots, n \quad (3)$$

step2.2: 如果 $|cluster| > \text{maxnum}$,对 cluster 中的文档用 CLARA 算法聚类,聚类结果加入 ClusterSet 中。

Step2.3: 将 cluster 从 ClusterSet 中删除,转 step2;

step3: 若 SampleBase 非空,转 step1,否则,结束。

1.2 基于范例推理

基于范例推理的过程即分类过程,如图2所示。先按2.1的方法将待分类文档 X 表示成特征向量,然后遍历范例库,计算出 X 与每个范例的相似度,选择相似度最大的 k 个范例 $\{S_1, S_2, \dots, S_k\}$ 。设这 k 个范例中有 n 个($n \leq k$)与待分类文档的相似度大于阈值 θ ,则根据这 n 个范例进行决策,计算与 X 相似度最大的类。设这 n 个范例中有 m 个 $\{S_{i_1}, \dots, S_{i_m}\}$ 属于类 C_i , X 与类 C_i 的相似度如下计算:

$$\text{sim}(X, C_i) = \sum_{p=1}^m \text{sim}(X, S_{p_i}) / m \quad (4)$$

对 n 个范例所属的所有类别计算其与待分类文档的相似度,设相似度最大的为类别 C 。如果满足 $\text{sim}(X, C) > \theta$,则可将 X 划分为类别 C 。设置阈值 θ 的目的是为了保证有足够近的最近邻,以免在与范例相似度都非常低的情况下把 X 分到一个相似度相对较大的类中。若没有满足以上条件的类别 C 时,则认为该文档用现有的分类器无法分类,需要对范例库进行修正。

1.3 范例库的修正

由于训练样本数量有限,分类器不可能对所有的文档都能正确分类,这就需要在使用的过程中通过学习的方法增强分类器的推理能力。当某个文档不能被正确分类时,本文先将该文档提交给用户,由用户对它进行手工分类,然后将它作为一个范例加入到正确类别的范例库中或定义新的类别来描述这类文档,同时当作一个训练样本保留在相应的训练样本库。对每类新增的范例计数,当新增范例在范例库中超过一定比例,则将该类的训练样本重新聚类生成范例。

2 实验结果及分析

2.1 测试数据及评价指标

为测试基于 CBR 的文本自动分类的性能,本文根据该方法实现了一个文本自动分类系统,以此为平台进行实验。文中的实验数据为中文自然语言处理开放平台上用于文本分类的语料库^[8],是由复旦大学计算机信息与技术系国际数据库中心自然语言处理小组提供的,该语料库共包括20个类别,共有训练文档9804篇,测试文档9833篇。由于有些类别的文档数比较少,实验时我们仅选择了其中数据超过500篇的6个类别: Agriculture、Art、Computer、Economy、Environment

和 Space。本实验用所有的训练文档来训练分类器,然后从每个类别随机选取 250 篇文章作为测试集。

本文采用召回率(公式 5)和准确率(公式 6)作为基本测试指标,并计算平均召回率(公式 7)和平均准确率(公式 8)评价系统的整体性能。

$$\text{召回率}(\text{recall}) = \frac{\text{分类的正确文档数}}{\text{应有文档数}} \quad (5)$$

$$\text{准确率}(\text{precision}) = \frac{\text{分类的正确文档数}}{\text{实际分类的文档数}} \quad (6)$$

$$\text{平均召回率} = \frac{\sum_{i=1}^n \text{类 } C_i \text{ 的召回率}}{n} \quad (7)$$

$$\text{平均准确率} = \frac{\sum_{i=1}^n \text{类 } C_i \text{ 的准确率}}{n} \quad (8)$$

2.2 分类性能实验结果及讨论

若对每个范例覆盖的训练样本数不加任何限制,则由于每个类里面的训练样本都比较相似,会出现大部分样本聚到一个范例的情况,此时,该范例包含了很多特征项,会导致该范例与待分类文档的相似度高于其他范例,从而影响分类结果。因此,我们设定了一个阈值 maxnum,表示每个范例覆盖的最大训练样本数。本文设计了三组实验,分别取 maxnum 为 100、80 和 50,实验结果如下:

表 1 maxnum 取 100、80、50 的召回率(%)

maxnum	Agriculture	Art	Computer	Economy	Environment	Space
100	92.80	90.80	97.60	94.80	86.80	43.60
80	96.40	99.60	97.60	93.60	65.20	70.00
50	96.80	90.80	97.60	94.40	84.00	43.20

表 2 maxnum 取 100、80、50 的准确率(%)

maxnum	Agriculture	Art	Computer	Economy	Environment	Space
100	92.06	99.56	64.55	84.04	88.57	96.46
80	79.54	98.03	75.31	91.05	97.02	94.09
50	88.97	99.56	64.72	85.82	89.74	96.43

表 1 和表 2 为三组实验中每类的召回率和准确率,其平均值如图 3 所示。从实验结果可以看出,当范例覆盖的最大训练样本数为 80 时,分类的效果最好。可见,分类系统的性能与 maxnum 并不是简单的线性关系,对不同的训练样本集,要通过实验来确定使系统性能最优的 maxnum 值。

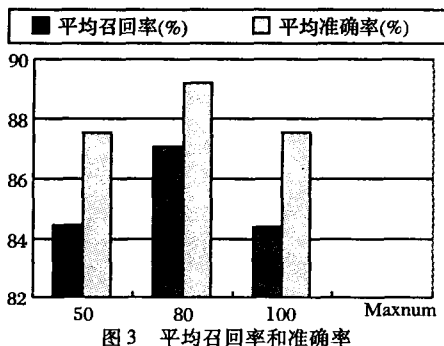


图 3 平均召回率和准确率

由文献[6]可知,KNN 方法的平均召回率为 83.29%,平均准确率为 85.12%,本文提出的基于范例的文本自动分类方法在 maxnum 取 80 时平均召回率和准确率分别为 87.07% 和 89.17%,可见,与 KNN 方法相比,本文提出的基于 CBR 的文本自动分类方法,通过把训练样本聚类成范例,使分类的召回率和准确率都有了提高。

2.3 时间复杂度分析

设训练样本库共有文档 m 篇,分别属于 k 个类别,特征项的维度为 n 维,KNN 算法分类的时间复杂度为 $O(km + nm)$ [6],本文将训练文档聚类为范例,则分类时间复杂度变为 $O(km' + nm')$,其中 m' 为范例库中范例的个数。由于每个范例覆盖了很多训练样本, m' 比 m 小得多;表 3 为各类的训练样本数和聚类后得到的范例数,从该表也可以看出,聚类后的范例数要远少于训练样本数。因此,根据上述时间复杂度的计算公式知,本方法在一定程度上克服了 KNN 方法分类速度慢的缺点,更能满足用户对实时性的需求。

表 3 各类的训练样本数和范例数

maxnum	Agriculture	Art	Computer	Economy	Environment	Space
100	55	45	47	47	38	36
80	55	53	71	47	63	36
50	91	77	516	723	410	40
训练样本数	1021	740	1357	1332	1217	640

3 结语

KNN 方法实现简单,分类召回率和准确率很高,在文本自动分类领域有着广泛的应用,但它在分类时要计算待分类文档与所有训练样本的相似度,时间复杂度很大。本文借鉴 KNN 的思想,将 CBR 技术与自动分类相结合,提出了一种基于 CBR 的文本自动分类方法。该方法先将训练样本聚类,把训练样本库转换为范例库,然后用 KNN 的思想分类。根据该方法实现了一个文本自动分类系统,并用中文自然语言处理开放平台上提供的用于文本分类的语料库对其测试,实验结果表明,该方法保留了 KNN 的优点,分类的平均召回率和准确率达到了 87.07% 和 89.17%;分析算法的时间复杂度可知,与 KNN 方法相比,该方法使分类速度有了很大的提高。综上所述,基于 CBR 的文本自动分类方法具有很好的实用价值。

文中采用向量空间模型来表示范例,而向量空间模型是一种文档表示的统计模型,它没有考虑文档上下文之间的语义关系,使得分类的召回率和准确率不高。因此,本文下一步的工作是,从语义方面对范例表示作进一步研究,以提高分类性能。

参考文献:

- [1] YANG Y, LIN X. A re-examination of text categorization methods [A]. The 22nd annual Int'l ACM SIGIR Conf. On Research and Development in Information Retrieval [C]. New York: ACM Press, 1999.
- [2] DEVIJVER P, KITTLER J. Pattern Recognition: A Statistical Approach [M]. Englewood Cliffs: Prentice Hall, 1982.
- [3] KUNCHEVA LI. Editing for the k-nearest neighbors rule by a genetic algorithms [J]. Pattern Recognition Letters, 1995, 16(8): 809 - 814.
- [4] KUNCHEVA LI. Fitness functions in editing KNN reference set by genetic algorithms [J]. Pattern Recognition, 1997, 30(6): 1041 - 1049.
- [5] MANTARAS RL, PLAZA E. Case-Based Reasoning: An overview [J]. AI Communications Journal, 1997, 10(1): 21 - 29.
- [6] 庞剑锋, 卜东波, 白硕. 基于向量空间模型的文本自动分类系统的研究与实现 [J]. 计算机应用研究, 2001, 18(9): 23 - 26.
- [7] KAUFMAN L, ROUSSEUW PJ. Finding Groups in Data, An Introduction to Cluster Analysis [M]. John Wiley & Sons, Brussels, Belgium, 1990.
- [8] http://www.nlp.org.cn/docs/download.php?doc_id=281 [EB/OL].