第33卷 第12期 计 算 机 工 程 2007年6月
Vol.33 No.12 Computer Engineering June 2007
·软件技术与数据库· 文章编号：1000—3428(2007)12—0096—03 文献标识码：A 中图分类号：TP181

一个对不带类别标记文本进行分类的方法

蒋志方[1]，祝翠玲[2]，吴 强[1]
(1. 山东大学计算机科学与技术学院，济南 250061；2. 山东经济学院信息管理学院，济南 250014)

摘 要：利用无监督聚类方法和朴素贝叶斯分类的特点，把UC获得的预分类结果作为朴素贝叶斯分类器的训练样本，将处在聚类结果中类属模糊区域的文本交给训练好的朴素贝叶斯分类器再进行分类，实现了对不带任何类别标记文本的准确分类，可得到较准确的分类结果。
关键词：文本分类；无监督文本聚类；朴素贝叶斯分类；欧氏距离

Method of Unlabeled Texts Classification

JIANG Zhifang[1], ZHU Cuiling[2], WU Qiang[1]
(1. School of Computer Science and Technology, Shandong University, Jinan 250061;
2. College of Information Management, Shandong Economic University, Jinan 250014)

【Abstract】 Using the specialty of the unsupervised clustering and the naïve Bayes classification, the paper gives a method that gains results of the text clusters and takes some of results as the training samples of the naïve Bayes classifier and let the trained naïve Bayes classifier reclassify those texts in illegible area of the clustering results. Consequently the method can classify the unlabeled text accurately and also can gain a better result of classification.
【Key words】 Text classification; Unsupervised text clustering; Naïve Bayes classification; Euclid distance

文本挖掘是对海量、异质、非结构化的文本数据源进行的涉及数据挖掘、自然语言处理、计算语言学、信息检索及分类、知识管理等的综合研究领域。文本分类常由训练过程和测试过程构成。在训练过程中，首先用训练文本训练分类器，得到最优的文本特征集合。在测试过程中，首先将测试文本用最优特征子集表示，再经分类器分类，得到测试文本所属的类别。文本聚类把一组文档按照相似性划分为若干类，属于同类的文档间的距离尽可能小而不同类文档间的距离则尽可能大[1]。

带标记训练样本的获得常由人工来完成的，随着文本数据的快速增长，这种方式的代价越来越高。故对文本的自动分类已成为组织和管理文本数据的关键技术，如何实现在非人工获得带标记训练样本和非人工预分类的情况下的准确文本分类，是分类方法研究的一个重要研究课题。

为实现对文本的自动分类，可采用聚类方法来产生训练样本和得到对文本分类的类别标记。本文提出了一个对不带任何类别标记的文本进行准确分类的方法。采用向量空间模型(vector space model，VSM)来表示文本，即将文本表示成在 n 维向量空间中的一个点。对在向量空间中的文本，指定聚类半径 R，利用无监督聚类(unsupervised clustering，UC)进行聚类，获得文本类别标记集合和聚类的正例中心和反例中心，然后把聚类结果中处在包含正例中心区域内的文本作为训练样本对朴素贝叶斯(naïve Bayes，NB) 分类器进行训练，最后将在聚类结果中处于类属模糊区的文本让训练好的NB 分类器再分类，可得到较准确的分类结果，提高分类精度。

1 无监督文本聚类算法

无监督文本聚类算法(unsupervised text clustering，UTC)是一种把UC算法[2~4]用于文本聚类的方法。在此算法中，通过指定聚类半径R，分别对每类文本进行聚类并获得文本聚类中心；然后，把文本聚类中心作为对文本的预分类的根据，即对任意给定的文本，计算其与各聚类中心的距离；离其最近的聚类中心所对应的类就是该文本的所属类。

1.1 算法描述

设有m个参加聚类的文本，一个文本由n个特征词来表示，记为 $X_i(A_{i1}, A_{i2},..., A_{in})$，形成$n$维向量空间中的一个点；记聚类结果为 $C_j, j=1,2,...,L$， L为到目前为止所形成的类的个数，C_j是一个属于同一个类的文本集合；O_j是类C_j的中心，也是n维向量空间中的一个点，记为 $O_j(B_{j1}, B_{j2},..., B_{jn})$，但$O_j$ 可能不是m个文本之一；记 k_j 为属于类C_j的文本的个数；记 $d(X_i, O_j)=\sqrt{\sum_{k=1}^{n}(A_{ik}-B_{jk})^2}$ 为第i个文本 X_i 和第j 个聚类中心 O_j的欧氏距离。

在对 m 个文本采用UTC算法进行聚类后，得到了L个类别 $C=\{C_1, C_2,..., C_L\}$，每个类C_j中的文本数 k_j 和所包含的文本 $\{X_{j,t} \mid t=1,2,..., k_j\}$，及其聚类中心$O_j$。把类$C_j$中包含的文本称为该类的正例集合，记为 Ω_j^+， m个文本中不属于类C_j的其他文本称为类C_j的反例集合，记为 Ω_j^-，其中文本个数为 $m-k_j$。

算法 1 无监督文本聚类算法(UTC)

输入 聚类半径 R；m 个 n 维文档集合 $Z=\{X_1, X_2,..., X_m\}, X_i \in R^n$ 为第 i 个文本；.

作者简介：蒋志方(1961－)，男，副教授，主研方向：数据挖掘与信息可视化；祝翠玲，硕士、助教；吴 强，博士生
收稿日期：2006-08-24 **E-mail：**zfjiang@sdu.edu.cn

输出 聚类结果C_j，聚类中心O_j ，C_j中所包含的文本集合。

步骤：

(1) $C_1 \leftarrow \{X_1\}$，$L \leftarrow 1$，$O_1 \leftarrow X_1$， $Z \leftarrow Z - \{X_1\}$

(2)若 $Z = \Phi$，则转至(7)。

(3)选择 $X_i \in Z$，从已有聚类中心中寻找与 X_i 最接近的中心O_j，即

$$O_j = \arg\min_{j}^{L}{}_{k=1} d(X_i, O_k)$$

(4)若 $d(X_i, O_j) < R$， 则将 X_i 加入类C_j，即

$C_j \leftarrow C_j \cup \{X_i\}$

调整C_j类的中心为

$$O_j \leftarrow \frac{k_j \times O_j + X_i}{k_j + 1}, k_j \leftarrow k_j + 1$$

然后转到(6)。

(5)否则 $L \leftarrow L+1, C_L \leftarrow \{X_i\}, O_L \leftarrow X_i$。

(6) $Z \leftarrow Z - \{X_i\}$，然后转到(2)。

(7)以类C_j的聚类中心O_j作为该类的正例中心 O_j^+，计算出每个类C_j的反例中心 O_j^-。

(8)对m个文本中的任意文本X，分别计算出其与类C_j的正例中心 O_j^+ 的欧氏距离和与反例中心 O_j^- 的欧氏距离：

$$d_{X,j}^+ = d(X, O_j^+)$$

$$d_{X,j}^- = d(X, O_j^-)$$

(9)求 X 与所有正例中心之间欧氏距离的最小值和 X 与所有反例中心之间欧氏距离的最小值。

$$d_X^+ = \min_{j=1}^{L} d_{X,j}^+$$

$$d_X^- = \min_{j=1}^{L} d_{X,j}^-$$

(10)决定文本 X 的类属：

若 $d_X^+ < d_X^-$ ，则 $X \in C_j$ ；若 $d_X^+ > d_X^-$ ，则 $X \notin C_j$ ；若 $d_X^+ = d_X^-$，则 X 类属不确定。

1.2 算法复杂度分析

算法从步骤(2)到步骤(3)为一循环过程。算法要对每个文本 $X_i (i = 1,2,...,m)$ 计算其与L个类的聚类中心的距离，所以该循环过程的时间复杂度是$O(L^2mn)$。在步骤(7)中计算反例中心的时间复杂度为$O(L^2)$，而在步骤(8)至步骤(9)的时间复杂度为 $O(2mnL)+O(mn)$。所以整个算法的时间复杂度是 $O(L^2mn)+ O(L^2)+ O(2mnL)+O(mn)=O(L^2mn)$。因此该算法具有较高的效率。

1.3 分类准确性分析

m个文本中的每一个X都可以用上述算法被归属于一个类C_j。事实上，对于任意给定的文本集合，在向量空间中，存在一个 $\varepsilon(\varepsilon > 0)$ 区域，其示意如图 1 所示。图中点划线所表示的区域就是ε区域，由欧氏距离公式和 $|d_X^+ - d_X^-| < \varepsilon$ 可知图中曲线为双曲线。

当 $|d_X^+ - d_X^-| > \varepsilon$ 时，文本 X 在此区域之外。当 $d_X^+ - d_X^- > \varepsilon$ 时，可以确定 $X \in C_j$，而当 $d_X^- - d_X^+ > \varepsilon$ 时，则可以确定 $X \notin C_j$。

当 $|d_X^+ - d_X^-| < \varepsilon$ 时，文本X在此区域内时，X与C_j的正例中心和反例中心距离相差不多，对其的归类就不一定准确。

可把ε作为一个决策阈值，来确定那些利用 UTC 方法可进行准确分类的文本，当 $|d_X^+ - d_X^-| > \varepsilon$ 时，可保证分类的准确性。

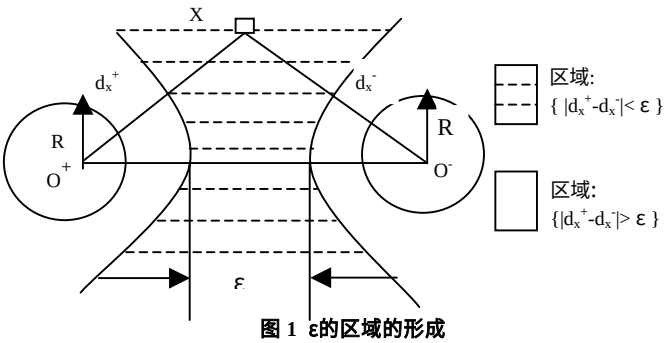

图 1 ε的区域的形成

2 朴素贝叶斯分类

2.1 分类方法

朴素贝叶斯分类(Naïve Bayes classfication，NBC)[1,5,6]是一种基于概率的分类法，它假设一个属性对给定类的影响独立于其他属性，即特征独立性假设。当假设成立时，与其他分类算法相比，NBC算法是最精确的[9]。

假设有L个类 $C_1, C_2, ..., C_L$。给定一个没有类标记的未知数据样本 $X(a_1, a_2, ..., a_n)$（a_i 表示第i个属性的值），NBC将预测X属于具有最高后验概率 $P(C_j \mid X)$ (条件X下)的类C_j，即将未知的样本分配给类C_j，当且仅当

$$P(C_j \mid X) > P(C_i \mid X), 1 \le i \le L, i \ne j \tag{1}$$

根据贝叶斯定理，有：

$$P(C_j \mid X) = \frac{P(X \mid C_j) P(C_j)}{P(X)} \tag{2}$$

为了降低计算 $P(X \mid C_j)$ 的开销，由 NBC 的特征独立性假设，有：

$$P(X \mid C_j) = \prod_{k=1}^{n} p(A_k = a_k \mid C_j) \tag{3}$$

其中，A_k表示样本X的第k个属性；a_k表示属性A_k的值，n表示属性的个数。因此式(2)可以改写为

$$P(C_j \mid X) = \frac{\prod_{k=1}^{n} p(A_k = a_k \mid C_j) P(C_j)}{P(X)} \tag{4}$$

式(4)中，$P(X)$ 是一个定值，由于 NBC 只需要比较 $P(C_j \mid X)$ 的大小，而无需计算 $P(C_j \mid X)$ 的值，因此对 $P(X)$ 可不予计算。

贝叶斯理论把在训练数据中得到的概率作为条件概率，因此式(4)中的一些概率可以直接从训练集获取。$P(C_j)$ 为任意一个样本属于类C_j的概率，若记类C_j中的样本个数为 $|C_j|$，所有样本总数为 $|S|$，则有：

$$P(C_j) = \frac{|C_j|}{|S|}$$

(1)如果A_k是离散变量，则有

$$P(A_k = a_k \mid C_j) = \frac{|C_j \mid A_k = a_k|}{|C_j|} \tag{5}$$

其中，$|C_j|A_k=a_k|$表示类C_j中属性A_k的值为a_k的样本数。

(2)如果A_k是连续值属性，则通常假定该属性服从高斯分布。因而，

$$P(A_k = a_k \mid C_j) = g(a_k, \mu_{c_j}, \sigma_{c_j}) = \frac{1}{\sqrt{2\pi}\sigma_{c_j}} e^{-\frac{(a_k - \mu_{c_j})^2}{2\sigma_{c_j}^2}} \tag{6}$$

其中，给定类C_j的训练样本属性A_k的值；$g(a_k, \mu_{c_j}, \sigma_{c_j})$ 是属性

A_k的高斯密度函数；μ_{c_j}和σ_{c_j}分别为平均值和标准差。

那么，对未知样本X分类，对每个类C_j，计算$P(X|C_j)P(C_j)$，样本被指派到C_j，当且仅当

$$P(X|C_j)P(C_j) > P(X|C_i)P(C_i), 1 \le i \le L, j \ne i \quad (7)$$

即X被指派到$P(X|C_j)P(C_j)$最大的类C_j。

2.2 特点分析

(1)在独立性假定成立时，可获得精度最优的分类效果；

(2)具有更小的出错率、更高的健壮性和效率；

(3)计算量较大。

3 无类别标记的文本的准确分类方法

3.1 方法的提出

无监督文本聚类的速度快但准确率较低，会出现不能准确判断其类别归属的模糊区域，而 NBC 在特征独立假设的前提下，具有分类准确度高，但计算量比较大的特点。可以只取出那些处于模糊区域的文本交给 NBC 来进行准确的分类。这样，将 UTC 方法和 NBC 方法结合起来，就有可能既保证有较快的分类速度，又有较高的分类准确率。由此本文提出一种基于无监督文本聚类和 NBC 的分类方法，能够对不带任何类别标记的文本进行准确的分类。

3.2 方法的描述

对于任意给定的文本集合，可以给定一个决策阈值$\varepsilon(\varepsilon > 0)$，首先对所有文本运用无监督文本聚类算法，产生一个聚类结果。在 1.2 节中，划分出了一个ε区域。处于ε区域之外，满足$d_X^- - d_X^+ > \varepsilon$的文本由 UTC 就能得到比较准确的分类，可以把这些文本作为 NB 分类器的训练样本，来训练 NB 分类器。而处在ε区域之内的文本，用 UTC 方法对其进行分类出错的概率较高，所以可以把该文本交给经过训练的 NB 分类器进行重新分类。

算法 2 对无类别标记的文本的准确分类算法(UNBTC)

输入 聚类半径 R，决策阈值ε，待分类的文本集合$Z = \{X_1, X_2, ..., X_m\}, X_i \in R^n$。

输出 聚类结果C_j，C_j中所包含的文本。

步骤：

(1)设定聚类半径 R，决策阈值ε。

(2)UTC算法对文本集合进行聚类，得到C_j(j=1,2,3,…,L)。

(3)对$|d_X^+ - d_X^-| > \varepsilon$区域中的文本，得到具有明确类属的文本的集合

(4)用$d_X^+ - d_X^- > \varepsilon$区域中的已有类别归属的文本对 NB 分类器进行训练

(5)用经过训练的 NBC 分类器对满足$|d_X^+ - d_X^-| < \varepsilon$的那些文本进行再分类。

(6)输出文本分类结果。

在本算法中，当待分类文本在文本向量空间中具有清晰类属时，应用了 UTC 算法的高效性与准确性，对处于向量空间中模糊区域的文本，则利用了 NBC 的分类错误率。

3.3 实验结果

在UCI知识发现数据库中的newsgroups数据集上对上述算法进行了试验，选取了 5 个分类中的 123 篇文档，选取了 276 个关键词对各个文档进行表示。表 1 是当聚类半径 R 取不同值时对 guns 类、mideast 类、religion 类的分类准确度(%)对比情况。

表 1 NBC 和 UTC 聚类准确率 (%)

accuracy	guns	mideast	religion
NBC	87.81	85.37	92.68
UTC(R=0.60)	77.24	65.04	82.93
UTC(R=0.50)	79.67	75.61	86.18
UTC(R=0.45)	83.74	82.11	88.61
UTC(R=0.40)	81.30	82.93	87.81

由表 1 可以看出，无论采用 NBC 还是采用 UTC，对于 religion 类的数据得到的结果准确率都高于其余两类。这说明在文本向量空间中该类距离其它类比较远，界限较清晰。对于这 3 个类别而言，NBC 的准确率均要高于 UTC，故 NBC 确有比较精确的区分能力。在 UTC 中，随着聚类半径 R 的不同，聚类结果也有不同的变化，当 R＝0.45 时达到较好的分类效果。运用该方法，当 R＝0.45 时，取不同的ε值时最后得到的结果如表 2 所示。

表 2 R＝0.45 时ε取不同值时的分类准确率 (%)

Class / ε	guns	mideast	religion
0.10	84.55	82.11	88.61
0.15	85.37	82.93	89.43
0.18	86.99	83.74	91.06
0.20	85.37	83.74	91.87

由表 2 中可以看出，这 3 个类别的分类正确率总体呈增长趋势，但也出现一些异常的情况，这可能是由多种情况引起的，例如：在计算权重时忽略了关键词在文中位置对其在文中重要性的贡献度；在利用 UTC 算法得到的结果对 NBC 器进行训练时选取的训练文本的数量以及准确度方面的提高方面等，还有待进一步改进。

4 结语

将 UTC 法与 NBC 法相结合是一种有效的分类方法，它实现了对不带任何类别标记的文本的准确分类，并解决了 UTC 准确率低和 NBC 方法计算量大、需要预分类的问题，充分利用了 UTC 分类速度快和 NBC 精度高的优点，获得了较高的训练速度和较高的识别准确率。

参考文献

1 Han J, Kamber M. 数据挖掘概念与技术[M]. 范 明, 孟小峰, 译. 北京: 机械工业出版社, 2001.

2 Eskin E, Arnold A, Prerau M, et al. A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data[M]//Applications of Data Mining in Computer Security. Boston: Kluwer Academic Publisher, 2002.

3 罗 敏, 王丽娜. 基于无监督聚类的入侵检测方法[J]. 电子学报, 2003, 31(11).

4 Doughtergy J. Supervised and Unsupervised Discretization of Continuous Features[C]//Proceedings of the 12th International Conference on Machine Learning. 1995.

5 Plangley W, Thompson K. An Analysis of Bayesian Classifiers[C]// Proc. of the 10th National Conf. on Artificial Intelligence. 1992.

6 Nigam K, McCallum A, Thrun S. Learning to Classify Text from Labeled and Unlabeled Documents[C]//Proceedings of the 15th National Conference on Artificial Intelligence. 1998.

7 McCallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification[C]//Proc. of the Workshop on Learning for Text Categoriazation. 1998.

8 宫秀军, 刘少辉, 史忠植. 一种增量贝叶斯分类模型[J]. 计算机学报, 2002, 25(6).