无标记训练样本的 Web 文本分类方法*)

刘丽珍[1] 宋瀚涛[2] 陆玉昌[3]

(首都师范大学信息工程学院 北京 100037)[1] (北京理工大学计算机系 北京 100081)[2]

(清华大学计算机系 北京 100084)[3]

摘 要 在文本分类中获得有类别标记训练样本的代价是很高昂的,本文针对这个问题对传统的模糊聚类方法进行改进,提出模糊划分聚类方法 FPCM,将聚类的无监督性和样本的先验知识结合起来,通过相似度度量聚类相关文本,取得比较客观的簇和少量标记文本,为监督学习找到分类依据,并结合朴素贝叶斯增量学习方式进行分类器的学习。本文进一步用估计分类误差损失的方法平衡选取候选样本,提高了分类准确率,实现了应用范围更加广泛的无标记文本分类学习模型。

关键词 Web 文本分类,模糊聚类,朴素贝叶斯

The Method of Web Text Classification of Using Non-labeled Training Sample

LIU Li-Zhen[1] SONG Han-Tao[2] LU Yu-Chang[3]

(Information Engineering College, Capital Normal University, Beijing 100037)[1]

(Department of Computer, Beijing Institute of Technology, Beijing 100081)[2]

(Department of Computer, Tsinghua University, Beijing 100084)[3]

Abstract Bayes learning theory is to obtain estimate of non-labeled samples by transcendental information and sample data. The application of text classification is to classify non-labeled texts by learning labeled class samples. But it is very difficult to obtain labeled training samples. In the paper the problem is analyzed in point of clustering view. The clustering is a non-supervised learning method, and has a character of independence on defined classes and labeled training samples. The thesis improve on tradition fuzzy clustering to bring forward Fuzzy Partition Clustering Method (FPCM). FPCM is a dynamic clustering method based on centroid technique. A few labeled texts are obtained to find classification foundation for supervised learning by fuzzy Partition clustering non-labeled Web texts. The sample's transcendental knowledge and clustering's non-supervisory are combined, and correlation texts are clustered by measuring similar degree. Naive Bayes augment learning style is further used to design and learn classifier. At the same time, classification precision is advanced using the way of selecting balance candidate samples after estimating the loss of classifying error. The model of text classifying using non-labeled training sample with more extensive application is realized.

Keywords Web text classification, Fuzzy clustering, Naïve Bayes

1 引言

Bayes 学习理论是利用先验信息和样本数据来获得对未知样本的估计,应用在文本分类中是通过对已标记类别样本的学习来分类未标记文本[1],但获得有类别标记训练样本的代价是很高昂的,本文从聚类的角度对这个问题进行了分析,提出了基于质心技术的动态模糊划分聚类方法 FPCM(Fuzzy Partition Clustering Method)。

分类和聚类的最大区别是:分类需要事先知道分类所依据的属性值,而聚类的方法是要找到这个分类的属性值;分类是根据应用的需要确定其类别,根据表示事物特征的数据识别其类别,但聚类是一种无监督学习,其类别不是人为指定的,而是分析数据的结果,通过比较数据的相似性和差异性,发现数据的内在特征及分布规律,从而获得对数据的理解和认识[2]。

聚类分析作为文本挖掘的一项关键技术,因为是无监督学习,不依赖于预先定义的类和带类别标记的训练实例,所以可以进行有监督分类的前期处理操作,也就是找出分类依据。另外,由于类别本身是人为定义的具有模糊性的概念,而各个特征属性和类别的关系也是模糊的,因此我们可以将模糊聚类理论和自动分类结合起来。

本文结合聚类的无监督性和样本的先验知识,用"物以类聚"的观点对无标记样本进行模糊划分聚类,通过一定的相似度度量,将相关文本归并,形成一定的先验信息,缩小搜索空间,然后用监督学习的分类方法训练分类器,构造出更加优越的分类学习模型。从实际的角度分析,聚类与分类的结合可以扩大分类适用范围,提高分类精度。

2 模糊划分聚类方法 FPCM

模糊聚类是对那些界限不分明的对象进行无监督分类,该方法利用建立在模糊关系基础上的相似关系,衡量事物之间的亲疏程度,并以此来实现分类。其分析的实质是依据一

*)基金项目:973 国家重点基础研究项目(G1998030414);北京市优秀人才专项经费资助项目(20042D0501604)。刘丽珍 副教授,博士,主要研究领域为数据仓储及知识发现;宋瀚涛 教授,博士生导师,主要从事多媒体与信息管理技术、网络通信技术的研究;陆玉昌 教授,从事数据集成和知识发现等的研究。

定的隶属度来确定其分类关系，得到样本属于各个类别的不确定性程度，体现样本类属的模糊性，建立了样本对于类别的不确定性描述，更加客观地反映了现实世界。

模糊划分聚类方法是一个基于质心技术的动态聚类方法[3]，通过优化一个准则函数把数据集划分成几个部分，一个划分表示一个聚簇，不断调整质心，通过迭代的重定位技术达到全局最优。

FPCM的评价聚类结果质量的准则函数是误差平方和准则，也称为目标函数。该方法对样本集进行初始划分模糊子集，设 N_i 是第 i 个聚类 Γ_i 中的样本数目，u_{ik} 表示样本 x_k 隶属于第 i 个聚类 Γ_i 的模糊隶属度，并保证：$\sum_{i=1}^{k} u_{ik}=1$，计算所有样本的模糊质心 v_i：$v_i=\frac{1}{N_i}\sum_{x\in\Gamma_i} x$，求所有样本与质心之间的误差平方和，形成目标函数或评价准则：

$$R=\sum_{i=1}^{c}\sum_{x\in\Gamma_i}\| x-v_i\|^2$$

R 是一个样本集和类别集的函数，度量了 c 个模糊聚类质心 $v_1,v_2,\cdots,v_c$ 所代表的 c 个模糊子集 $\Gamma_1,\Gamma_2,\cdots,\Gamma_c$ 产生的总误差平方，最后通过迭代运算找出目标函数的极小值，聚类得出最优结果。

在取得样本集的初始划分类以前，通常我们要先选取一些具有代表意义的点作为聚类的初始质心，才能将其余的点以某种方式划分到以这些初始点为核心的初始类中，这种简单快速的初始分类方法，可以使后面的整个聚类过程速度加快。

设：$X=\{x_1,x_2,\cdots,x_n\}$ 为文本集合，x_k 为 X 中样本，$i=1,2,\cdots,n$；每个样本 x_k 都有特征向量 $p(x_k)=(x_{k1},x_{k2},\cdots,x_{ks})$，其中 $x_{kj}(1\leqslant j\leqslant s)$ 是 x_k 的第 j 个属性值，$p(x_k)$ 为 x_k 的特征向量。模糊聚类分析就是将文本集合 X 中的 n 个样本所对应的特征向量进行分析，根据相似性的类属划分准则 R 将样本划分到 c 个模糊子集 $\Gamma_1,\Gamma_2,\cdots,\Gamma_c$，其中样本隶属函数 μ_{ik} 表示样本 x_k 隶属于第 i 个聚类的模糊隶属度，它的取值在[0,1]区间，supp 是模糊集合的支撑集，满足条件为[4]：

$$\bigcup_{i=1}^{n} SUPP(\Gamma_i)=X;\mu_{ik}\in[0,1];$$

$$\sum_{i=1}^{n}\mu_{ik}=1,\forall k;0<\sum_{i=1}^{n}\mu_{ik}<n,\forall i$$

模糊划分聚类方法 FPCM 比传统的模糊 C-均值聚类方法[5]的性能要好，传统的模糊 C-均值聚类每调整一个样本的类别就要重新计算一次各类样本的均值，而 FPCM 算法使用了成批样本的修正方法。另外，在本文的应用中，由于对聚类在精度上没有很高的要求，只要能通过聚类为后面的有监督学习提供有标记的训练样本即可，在聚类过程中不进行合并和裂变，因此 FPCM 方法比模糊 C-均值聚类的变种 ISODATA 算法的计算复杂度小，而且速度也比 ISODATA 快。FPCM 方法的计算复杂度是 $O(cnt)$，其中，n 是所有对象数目，c 是簇的数目，t 是算法迭代次数，k、$t\ll n$。

Web 文本的标题、段首、段尾出现的词汇对整篇文本的内容起到很重要的作用，通过标注可以很方便地得出词汇对文本内容的贡献程度，但它们都不能十分准确地加以定义，包含了很大的模糊性，而模糊划分聚类方法 FPCM 在处理数据相似性系数时更精确，聚类结果的解释更易于理解；再者，FPCM 对球状簇有很好的分类效果，而 Web 文本算法经过特征抽取后，在多维空间的分布恰好呈球形分布；另外，FPCM 计算效率高，该聚类方法可以随数据的大小线性地扩展，在数据增加的情况下具有良好的可伸缩性。模糊划分聚类算法 FPCM 在本文的使用中，对迭代次数要求不是很高，只要算法开始收敛即可，在聚类精度和迭代次数上放得比较宽可以体现出聚类速度较快的优势，本文使用该方法无论在理论上还是在应用上都是较好的选择。

3 模糊划分聚类 FPCM 与 Naïve Bayes 增量学习相结合的 Web 文本分类方法

在 Web 文本分类中，Naïve Bayes 模型使用最多的操作就是利用当前少数带有类别标记的训练样本对未标记文本进行分类，并且对增加训练样本后的模型进行学习和参数修正。新的训练样本的加入是综合先验信息和样本信息后形成样本知识(后验知识)的，经过分类后将其加入到训练集。如果要对形成的新模式重新学习和修正参数，必然导致很高的复杂性，但 Naïve Bayes 模型具有增量(Augment)学习的特性[6]，只需改变与变化相关的项的估计，所以大大简化了问题的复杂度，缩小了操作规模。

在监督方式下的分类学习中认为训练样本都是独立同分布的，为了提高模型的分类性能，引入一定的方法平衡候选学习样本的偏颇，均衡标记文本和未标记文本的数据分布是很有必要的[7]。

我们通常对候选样本的选取是根据分类误差损失最小，从有助于提高分类精度的角度出发，但这种有目的的选取分类样本，常常会导致最后入选内容范围涉及面很广或内容比较稀奇的文本进入分类器，而先入选分类的文本大都带有某种偏好，这样势必造成信息分布的偏斜性，使分类器缺乏对多类问题处理的有效性，影响分类器的分类性能。对于多数的实际分类文本大都存在上面的选取问题，所以本文利用分类损失误差来平衡分类样本选取的偏颇。

通过对模糊划分聚类 FPCM 和 Naïve Bayes 增量学习特性的研究，将二者结合起来，用于无标记训练样本的文本分类。这种半监督的分类方法运用无监督性和先验信息各自的特点和结合优势，通过相似度度量聚类相关文本，取得比较客观的簇和少量标记文本，为监督学习找到分类依据，利用 Naïve Bayes 增量方式对少量标记训练样本的高效率学习，再加上对候选分类样本的平衡选取，构造出了更加优越的分类学习模型。

模糊划分聚类 FCPM 结合 Naïve Bayes 增量学习的分类算法如下：

Step1：对无标记文本集 X 进行模糊划分聚类

Step2：将距模糊划分子集质心最近的样本作为标记文本放入训练集 U_A：

$\forall x_k,\exists v_i$ if $h_{ik}<\omega$ $U_A=U_A+[x_k]$，其余的放入集合 U_B

Step3：根据这些已标注类别的训练集 U_A 进行未标注文本的增量分类学习 $i\leftarrow 1$

Step4：$x\leftarrow x_i$，$x\in U_B$ If($U_B=\phi$) goto step8

Step5：求最小损失 minl

Step6：$D\leftarrow D+\{(x_{\min k},c_j)\}+\{(x_{\max k},c'_j)\}$

Step7：i++；goto step 4

Step8：输出分类器 C

4 实验与分析

实验数据来源于从 Internet 下载的 4800 篇关于计算机

(下转第 211 页)

的。

参 考 文 献

1 Okutomi M, Kanade T. A multiple-baseline stereo. In: Proceeding IEEE Conference on Computer Vision and Pattern Recognition, 1991. 63~69

2 Kanade T, Okutomi M. A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(9): 920~932

3 Fusiello A, Roberto V, Trucco E. Efficient stereo with multiple windowing. In: Proc. IEEE Conf. on Computer Vison and Pattern Recognition, 1997. 858~863

4 Geiger D, Ladendorf B, Yuille A. Occlusions and binocular stereo. In: Proc. European Conf. on computer Vison, 1992. 425~433

5 Veksler O. Stereo matching by compact windows via minimum ratio cycle. In: Proceeding International Conference on Computer Vision(ICCV), 2001, 1: 540~547

(上接第 201 页)

(800 篇)、教育(800 篇)、法律(800 篇)、艺术(800 篇)、体育(800 篇)和军事(800 篇)的六个类别的中文网页,本文在实验中分别用模糊 C-均值聚类算法、ISODATA 算法和 FPCM 算法进行聚类,迭代次数分别为 5、8 和 3,然后用 Naïve Bayes 增量学习方法分类,其分类的准确率如图 1 所示。

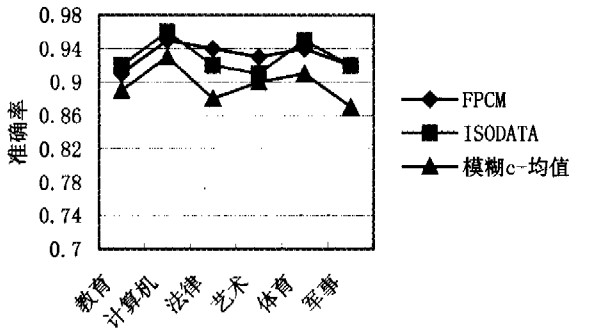

图 1 分别用 FPCM、ISODATA 和模糊 C-均值三种聚类方法结合 Naïve Bayes 增量学习进行无标记 Web 文本分类

从实验中我们得出,FPCM 和 ISODATA 聚类结合 Naïve Bayes 增量分类的结果准确率比较相近,模糊 C-均值聚类结合 Naïve Bayes 增量分类准确率比前两个方法要差。在实验中我们还观察到在整个分类的过程中,采用 FPCM 先聚类的分类速度最快,其次是模糊 C-均值, ISODATA 的速度相对较慢。在本文对聚类的应用精度要求不是很高的情况下,FPCM 无疑是最好的选择,因为 FPCM 比模糊 C 均值方法的精度高,比 ISODATA 的速度快。因此,在无标记文本分类中,用 FPCM 聚类作为 Naïve Bayes 分类的前期标记工作,能达到比较满意的分类结果。

以下我们用 FPCM 结合 Naïve Bayes 增量学习分类的结果与模糊 C-均值聚类、ISODATA 聚类的结果进行比较,如图 2 所示。

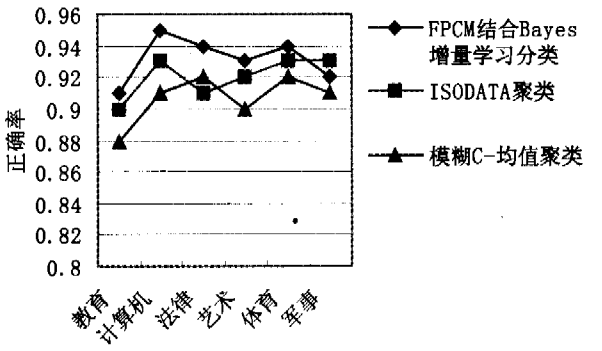

图 2 FPCM 结合 Bayes 增量学习分类结果和 ISODATA 聚类、模糊 C-均值聚类结果的比较

表 1 FPCM 与 Naïve Bayes 增量学习分类相结合的实验结果

	教育	计算机	法律	艺术	体育	军事
TP	705	752	758	762	750	758
FN	69	38	43	57	46	61
FP	95	48	42	38	50	42
Pre.	91%	95.3%	94.6%	93%	94.1%	92.5%
Rec.	88.1%	94%	94.8%	95.2%	93.6%	94.8%

从实验中,模糊划分聚类 FPCM 的迭代次数 t 为 3,控制模糊聚类参数 $m=2$,停止阈值 $0.4\leqslant\lambda\leqslant0.6$,$\lambda$ 越小,聚类越细,模糊度越小,聚类精度越高,且所需要的时间也就越长,所以在本文中我们没有将 λ 选得很小,尽量减少计算时间。通过实验我们观察到 FPCM 与 Naïve Bayes 增量学习相结合的分类方法,比单纯的两种模糊聚类的分类准确度高,分类性能好。表 1 给出了 FPCM 与 Naïve Bayes 增量学习分类相结合的实验结果。

结束语 在 Web 文本分类中,获得已标记训练样本集的代价是很高昂的。本章针对无标记训练样本的文本分类问题进行了研究,提出了模糊划分聚类 FPCM 与 Naïve Bayes 增量学习相结合的分类方法。

本文利用无监督聚类不依赖于预先定义的类和带类别标记的训练实例的特点,对模糊聚类进行改进,提出了模糊划分聚类方法 FPCM。使用该方法通过对 Web 无标记文本的模糊聚类,得到少量标记文本,从而为监督学习找到了分类依据,与 Naïve Bayes 增量学习方法的结合,提高了文本分类精度,对候选样本的平衡选取,进一步增强了分类器的性能。对于这个模型,我们进一步的工作是要解决经验选取初始点的敏感问题,以及尽量减小由于数据的分散性所带来的对孤立点的敏感程度,减少人工的干预,同时又不能增加太多的计算复杂度,在不同规模的数据集上测试该方法的有效性,以构造更高性能的分类器。

参 考 文 献

1 Linoff G S, J. a. Berry M. Mining the web, America, 2001, 348

2 Mena J. Data Mining your website. America, 2000, 368

3 Wang Shi, Gao Wen. Web data mining. Computer Science, 2000, 27(4): 237~240

4 Hutter M. Distribution of Mutual Information. In: Proc. of the 14th Intl. Conf. on Neural Information Processing Systems, NIPS-2001

5 边肇祺,张学工,等编著,模式识别(第二版),清华大学出版社,2000

6 Keogh E J, et al. Learning Augmented Bayesian Classifiers: A Comparison of Distribution-based and Classification-based Approache, 2002 http://citeseer.nj.nec.com/context.

7 宫秀军,等. 主动贝叶斯网络分类器. 计算机研究与发展, 2002, 5: 574~579