文本分类中粗分类数据噪声修正的网络算法 [1)]

宣照国　党延忠

（大连理工大学系统工程研究所，大连 116023）

摘　要：在文本分类的实际应用中经常使用粗略分类的数据来训练分类器，但是这种数据中经常会包含类别标记有误的数据，这些数据对文本分类结果的精度会造成不良影响。本文针对这个问题提出了一种噪声修正算法，首先建立文档关联网络，把文档上标记的类别作为在网络上划分的集团结构，并用模块度衡量集团结构的质量，通过优化模块度指标把噪声数据调整到合适的类别中，从而提高数据质量。实验结果表明，本文所提算法能够有效修正粗分类数据中的噪声，且有较高的有效性和鲁棒性。该算法可以用于文本分类训练数据的预处理，或作为辅助技术用于文献库建设等工作。

关键词：噪声数据修正；模块度优化；文本分类；集团结构

Network-based Noise Revision Algorithm in Text Categorization

Xuan Zhaoguo　Dang Yanzhong

(*Institute of Systems Engineering, Dalian University of Technology, Dalian 116023*)

Abstract: Training data is necessary to train the classifiers in Text Categorization. In fact, there are always some documents distributed to a wrong category in training text corpus, which are named noise texts. If we use noise texts in text mining applications directly, the efficiency of the text mining will be influenced. This paper proposes a revision algorithm for noise texts based on network. Firstly, document-similarity network (DSN) is constructed. The categories constitute the corresponding community structure in the network, and modularity is used to evaluate the quality of the categories. The noise texts can be revised through modularity optimization. The experimental results indicate the efficiency and robustness of the algorithm. This algorithm can be used in the preprocessing of text mining or taxonomy building.

Keywords: noise texts revision; modularity optimization; text categorization; community structure

1 引言

文本分类是文本处理的一项基本工作，提高文本分类精度已成为一项重要而迫切的研究课题。文本分类技术可以分析被分类文档的特征，并与各类别中文档所具有的共同特征进行比较，将被分类文档归为特征最接近的一类中并赋予相应的类别。文本分类是有指导的自动处理技术，分两个阶段进行。首先通过带有类别标签的训练文档集来构造分类器；然后使用构造出来的分类器对新文档进行分类。第一阶段中构造的分类器的质量，对第二阶段的分类精度具有直接的影响。但在实际应用中，无法预先提供类别准确的、

1)　国家自然科学基金重点项目(70431001),国家自然科学基金重大国际合作项目(70620140115)，国家自然科学基金资助项目(70271046,70301009)

作者简介：宣照国，男，1971 年生，博士生，主要研究方向：中文信息处理、文本知识发现。党延忠，男，1954 年生，博士生导师，教授，主要研究方向：知识科学、知识管理、系统开发与集成。

全面的精分类训练文档数据，往往直接使用一些粗略分类的数据，即粗分类数据作为训练文档，比如按照预先定义的栏目类别整理的网页数据、期刊上划分为各个栏目的论文、以及科研资助机构中标记着学科代码的立项建议书等。这些数据中具备了类别特征，即每个类别下的数据具有一定的共性特征，但是各个类别下的成员文档集合中，经常会存在一定数量的类别标记错误的文档，即文档内容与标记的类别不符。本文把这种类别标记错误的文档称为噪声数据。如果用这种含有噪声数据的粗分类文档数据训练分类器，必将对分类结果的准确性造成不良影响。

目前，很多研究人员致力于研究提高分类结果准确率的算法，比如通过特征筛选[1,2]、精选训练文档[3,4]、以及多种算法结合[5,6]等方法，来提高分类器的性能，而对提高训练数据质量的数据整理算法重视不够，研究也较少。本文提出了一种针对粗分类文档中噪声数据的修正算法，把类别标记错误的文档放回正确的类别中，获得精分类的训练数据，从而提高分类器的性能。其基本思路是：通过计算文档之间的相关度，构建文档关联网络模型，进而借鉴复杂网络理论中社区集团发现方法及其结果评价算法，对类别划分质量进行评价，并通过优化类别划分质量来修正文档集合中的噪声数据，从而获得分类准确的训练数据。

2 噪声数据对分类结果的影响

为了说明噪声数据对分类结果的影响，本文进行如下实验。

选择中文文本分类语料[7]作为实验数据，其中共包含 2815 篇文档，具有 10 个类别。根据传统的向量空间模型（Vector Space Model，VSM），文档的内容被表示为特征空间中的加权特征向量 $d_i=(t_1, \omega_{i1}; t_2, \omega_{i2}; \cdots; t_n, \omega_{in})$，$t_k$ 表示特征词条，ω_{ik} 表示词条 t_k 在文档 d_i 中的权值。ω_{ik} 使用 $tf\text{-}idf$ 计算，即：

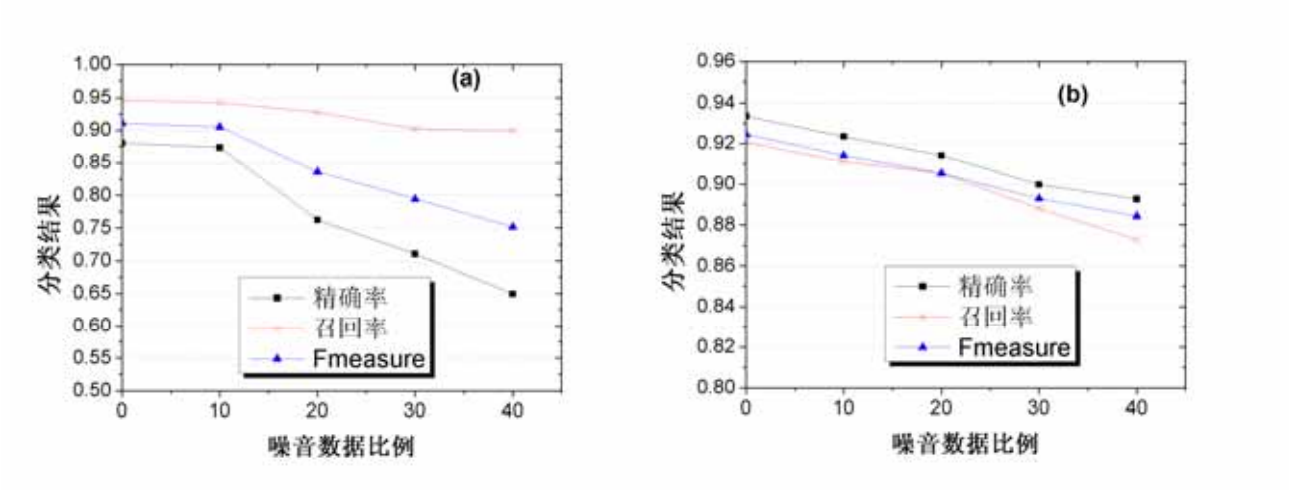
$$\omega_{ik}=\frac{tf_{ik}\log(n/n_k+0.01)}{\sqrt{\sum_{j=1}^{m}tf_{ik}{}^2(\log(n/n_k+0.01))^2}} \tag{1}$$

其中，tf_{ik} 表示词条 t_k 在文档 d_i 中出现的次数，n 表示文档集合 D 中的文档总数，n_k 表示集合 D 中包含特征词 t_k 的文档数。把文档表示为特征向量之后，使用夹角余弦公式来计算两篇文档之间的相似性。对于文档 d_i 和 d_j，它们之间的相似度 s_{ij} 的计算方法，如公式(2)所示：

$$s_{ij}=\frac{\sum_{k=1}^{n}\omega_{ik}\cdot\omega_{jk}}{\sqrt{\sum_{k=1}^{n}\omega_{ik}{}^2}\sqrt{\sum_{k=1}^{n}\omega_{jk}{}^2}}=\sum_{k=1}^{n}\omega_{ik}\cdot\omega_{jk} \tag{2}$$

我们把实验数据分为 5 组，其中 4 组作为训练文档集，另外 1 组作为测试文档集，分别使用了 kNN 算法和 Recchio 算法进行分类测试。首先直接使用训练文档来训练分类器并测试分类结果，然后逐步在训练文档集中增加噪声数据，通过对比分类结果来发现噪声数据对分类器性能的影响。在训练文档集中加入噪声的方法是：从每个类别的训练文档中随机选择出一定比例的文档，再把每个被选择出来的文档加入到另外的任一类别的训练文档集中，作为噪声数据。在实验中，逐步增加噪声数据的比例，测试不同比例的噪声数据对分类结果的影响。

使用精确率、召回率和 Fmeasure 衡量分类结果[8,9]。分别使用 kNN 算法和 Rocchio 算法进行分类实验，实验结果如图 1 所示。



(a)kNN分类结果(k=5)；(b) Rocchio分类结果

图1.噪声数据对分类结果的影响分析

从图1的实验结果可以看出，随着训练数据中噪声比例的增加，分类结果的质量逐步下降。其中kNN分类结果受噪声的影响尤其明显，当噪声比例为20%时，分类结果的准确率仅为75%。Rocchio分类结果的准确率也随着噪声比例的增加而下降。实验说明，训练文档中的噪声对文本分类结果具有严重的不良影响，因此有必要在分类之前对构造分类器的训练文档中的噪声数据进行修正，才能构造出高质量的分类器，从而保证分类结果的精度。

3 文档关联网络

3.1 网络的构建及集团结构

本文通过计算文档之间的相似度构建文档关联网络(Document-Relation Network, DRN)。在DRN中共包含了n个节点（$d_1, d_2, \cdots, d_n$），每个节点代表一个文档，节点之间的边代表两个文档是相似的，并把文档之间的相似度s_{ij}作为边的权。DRN可以用矩阵$S=[s_{ij}]_{nn}$来描述，称为文档关联矩阵。这个矩阵是对称且反自反的，满足$s_{ij}=s_{ji}$，$i, j=\{1,..,n\}$，且$s_{ii}=0, i=\{1..n\}$，即DRN网络不考虑文档自身的相似性。

这个文档关联网络DRN具有许多复杂网络的特性，本文提出的算法利用了其中的集团结构，其他特性不在此赘述。DRN中的集团是DRN中性质相近节点组成的类，同一个集团（类）中的节点之间连接比较紧密，而不同集团（类）中节点之间的连接比较松散[10,11]。

本文把粗分类作为DRN上的一种集团结构划分方案，每个集团就是一个粗分类的类。n个文档分为m个类($c_1, c_2, \ldots, c_m$)，每一个文档都被标记为m个类中的一个类。在DRN中共包含了m集团，每个集团代表一个类，两个集团之间的所有边上的权反映了两个集团（即两个类）之间的关联状况。以集团（类）作为节点，集团(类)之间的关联作为边，关联强度作为边的权值，这样就构成一个新的网络“集团网络”(Group-Relation Network, GRN)。本文使用矩阵$E=\{e_{ij}\}_{mm}$来描述集团网络，称之为集团矩阵。DRN是基础网络，GRN是建立在DRN上的集团（类）网络，根据文档关联矩阵S可以定义集团矩阵E，其元素e_{lk}定义如下：

$$e_{lk} = e_{kl} = \frac{1}{2*tot} \sum_{d_i \in c_l, d_j \in c_k} s_{ij} \tag{3}$$

其中，tot为网络DRN中所有边的权值总和，即：

$$tot = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} \tag{4}$$

矩阵E中对角线上的元素e_{kk}为集团c_k内部所有边的权值在网络中所占的比例，即：

$$e_{kk} = \frac{1}{2*tot} \sum_{d_i \in c_k, d_j \in c_k} s_{ij} \tag{5}$$

根据矩阵E的构建过程可以得出，$\sum_{l=1}^{m}\sum_{k=1}^{m} e_{lk} = 1$。

本文通过对各个集团的成员文档进行调整来修正噪声数据，使得网络上的集团结构划分得趋于合理。

3.2 集团结构的评价指标

Newman提出了评价拓扑网络中集团结构划分质量的指标，称为模块度(Modularity，记为Q)[12]。在拓扑网络中，根据网络集团之间边的数量来计算模块度。本文认为：在加权网络中，边的权值比边的数量更能够反映集团结构划分的质量。因此，本文使用边的权值来计算加权网络上的模块度(记为Q^{ω})。加权网络的模块度Q^{ω}的计算方法如下：

$$Q^{w} = \sum_{l=1}^{m} (e_{ll} - a_l^{\,2}) \tag{6}$$

其中，a_l为矩阵E中第l行元素的和，即：

$$a_l = \sum_{k=1}^{m} e_{lk} \tag{7}$$

当网络中所有边的权值为1时，Q^{ω}和Q的计算结果相等。模块度(Q^{ω})这一指标综合评估了集团内和集团间节点的连接强度，从网络结构角度衡量了集团结构划分的质量。如果权值大的边连接的节点被划分到同一类别中，那么模块度就大；反之，如果权值大的边连接的节点被划分到不同类别中，模块度就小。为考察噪声数据对模块度的影响，本文计算了不同比例噪声数据的模块度，结果如表1所示。

表1．噪声对模块度指标的影响

噪声数据比例	Q^{ω}
0%	0.6981
10%	0.5548
20%	0.4342
30%	0.3081
40%	0.2092
50%	0.1356
60%	0.0777
70%	0.0299

从表1中的计算结果可见，模块度随着噪声数据含量的增加而下降，反之模块度Q^{ω}越大，说明噪声数据的含量越少。因此模块度能够反映出集团结构，即类别划分的质量。如果能找到一种算法，把模块度作为优化指标使其不断增大，那么就可以减少粗分类文档中的噪声数据。这就是本文提出的“网络算法”的出发点。

4 噪声数据修正算法

为了检验一个文档是否为噪声数据，需要把文档从其所在的当前类别临时改成其他的类别。本文把文档d_i当前所在类别称为d_i的标记类别，记为Tag_i；把临时改成的类别称为d_i的测试类别，记为$Test_i$。

把一个文档d_i从标记类别Tag_i调整到测试类别$Test_i$，在文档关联网络DRN上，各个节点之间的边以及边的权值都没有改变，但集团Tag_i和$Test_i$的成员发生了变化。同时，与d_i相连接的其他文档d_k所在的集团Tag_k，与集团Tag_i、$Test_i$的关联关系也发生了变化。为了清楚地描述出这些变化，我们定义了文档与类别的隶属度。

定义：文档对类的隶属度。把文档d_i与类别c_j的所有成员之间边的权值在网络中所占的比例，称为文档d_i对于类别c_j的隶属度，记为Deg_{ij}，即：

$$Deg_{ij} = \frac{1}{tot}\sum_{d_k \in c_j} s_{ik} \tag{8}$$

由此可以构建隶属度矩阵$A=(Deg_{ij})_{mn}$，m，n分别表示集团（类）的个数和文档个数。

当一篇文档d_i所属的类别发生变化，矩阵A中的元素的值也会相应地发生改变。当d_i从类别x调整到类别y时，与文档d_i邻接的其他文档d_j对应的行的元素调整如下：

$$Deg_{jx}=Deg_{jx}-s_{ij}，\ Deg_{jy}=Deg_{jy}+s_{ij} \tag{9}$$

文档对于类别的隶属度反映了文档与类别之间的关联强度，隶属度越大，说明文档越倾向于归属于这个类别。调整一个文档的类别，与该文档邻接的其他文档的隶属度也相应改变，因此修正数据是一个动态调整过程。把文档d_i从标记类别Tag_i调整到测试类别$Test_i$中，设$x=Tag_i$，$y=Test_i$，调整后矩阵E的元素改变为：

$$e_{xx}=e_{xx}-Deg_{ix}，\quad e_{yy}=e_{yy}+Deg_{iy}，\ e_{xy}=e_{yx}=e_{xy}+(Deg_{ix}-Deg_{iy})/2 \tag{10}$$

而对于其他类别$z(z\neq x, z\neq y)$，有：

$$e_{xz}=e_{zx}=e_{xz}-Deg_{iz}/2，\ e_{yz}=e_{zy}=e_{yz}+Deg_{iz}/2 \tag{11}$$

可以使用公式（6）计算出调整前后的模块度，如果模块度增大，说明调整是合理的。

本文根据文档对于类别的隶属度来判断是否需要对其修正。设c_i^*是文档d_i的隶属度最大的类别，即$c_i^* = \arg(\max_j Deg_{ij})$，如果$c_i^*=Tag_i$，则不对其进行调整。否则，将$d_i$作为需调整的候选文档，并计算把$d_i$从类别$Tag_i$调整到类别$c_i^*$后的模块度$Q''(i)$。比较所有需调整的候选文档的$Q''$，选择具有最大$Q''$的文档$d_t$，把$d_t$从类别$Tag_t$调整到类别$c_t^*$中，同时更新隶属度矩阵A和集团矩阵E。如此反复，直到模块度不再增大为止。算法流程如下：

输入：DRN的节点集合$D=\{d_i\}$，边权集合$\{s_{ij}\}$，各个节点的原始类别$\{Tag_i\}$；

输出：各个节点的调整后的类别$\{T_i\}$；

1. 初始化：建立矩阵E和矩阵A；计算初始模块度Q'；设置$T_i:=Tag_i$；
2. DO WHILE (true)
3. BEGIN
4. FOR EACH $d_i \in D$ DO
5. BEGIN
6. $c_i^* := \arg(\max_{j=1}^{m} Deg_{ij})$；//获取$d_i$的隶属度最大的类别
7. IF $c_i^* = T_i$ THEN
8. $Q''(i) := Q'$；
9. ELSE
10. $Q''(i)$:=(把d_i从T_i调整到c_i^*后的模块度)；
11 END FOR

12. $t := \arg(\max_{i=1}^{n}(Q^w(i)))$ ；// d_t调整后的模块度最大
13. IF $Q^w(t)>Q^w$ THEN
14. $T_t := c_t^*$；
15. $Q^w := Q^w(t)$；
16. 使用公式(9)更新矩阵A；
17. 使用公式(10)和(11)更新矩阵E；
18. ELSE
19. 算法终止，输出修正结果$\{T_i\}$；
20. END WHILE

算法每个时间步对一个文档d_i进行调整测试，首先在矩阵A的第i列查找d_i的隶属度最大的类别c_i^*，复杂度为$O(m)$；然后计算从把d_i从当前类别T_i调整到类别c_i^*的模块度：先根据矩阵A调整矩阵E中元素的值，计算复杂度为$O(2m)$；再根据调整后的矩阵E的数据计算模块度，计算复杂度为$O(m^2)$。因此，每个时间步的计算复杂度为$O(m+2m+m^2)$，即$O(m^2)$。每次调整要测试所有文档，复杂度为$O(nm^2)$。假设经过T次调整得到最大的模块度，则计算复杂度为$O(Tnm^2)$。这里假设每个文档都可能需要进行类别调整测试，实际情况下，随着算法的执行，越来越多的文档都被标记为合适的类别（即$T_i=c_i^*$），不需要进行类别调整测试，因此算法实际计算复杂度要远远低于$O(Tnm^2)$。

5 实验结果分析

使用中文文本分类语料[7]作为实验数据。数据集中共包含10个类别，共2815篇文档。把语料库中每个类别的文档平均分成2组，分别作为正确数据和噪声数据。把噪声数据随机分配到正确数据的其他任一类别中，这样生成了含有50%噪声比例的粗分类数据。生成数据中有1409篇文档类别标记是正确的，另外1406篇文档的类别标记是错误的。使用本文提出的算法对这些粗分类数据进行噪声修正处理，修正过程中的模块度和正确率变化曲线如图2所示。

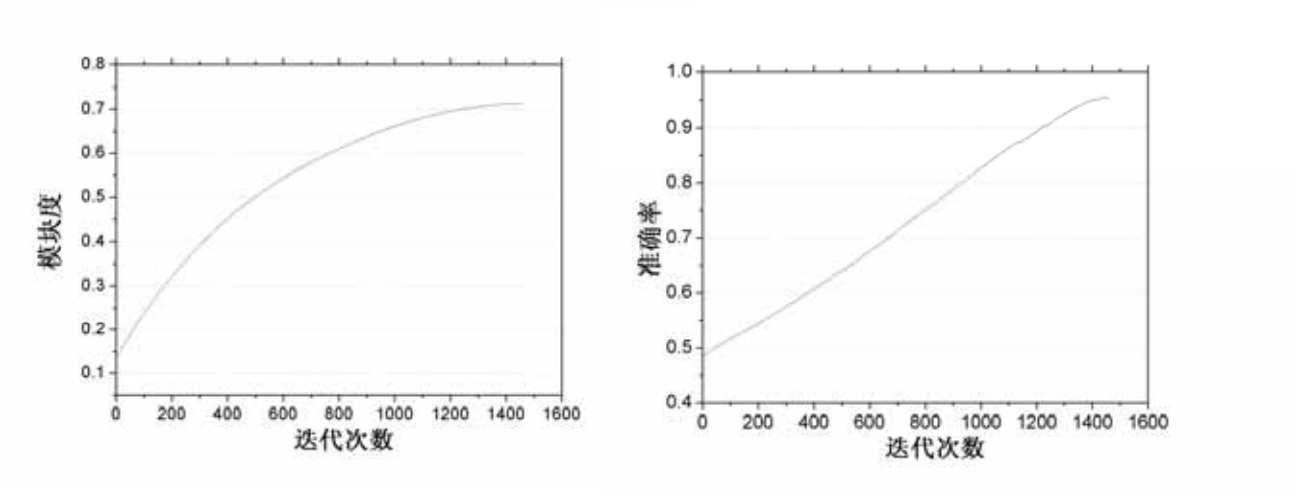


图2：修正过程中模块度和准确率的变化曲线

通过图2可以看出，随着算法进行修正，模块度和准确率同步增加。算法结束后，模块度达到0.7134，准确率达到0.9346。算法每次迭代修正1个数据，对于包含1406个噪声的实验数据，算法总共迭代了1463次。为了进一步验证算法效率，我们生成了包含不同噪声数据比例的实验数据，使用本文提出的算法对这些粗分类数据进行噪声修正处理，修正过程中的模块度变化曲线如图3所示。

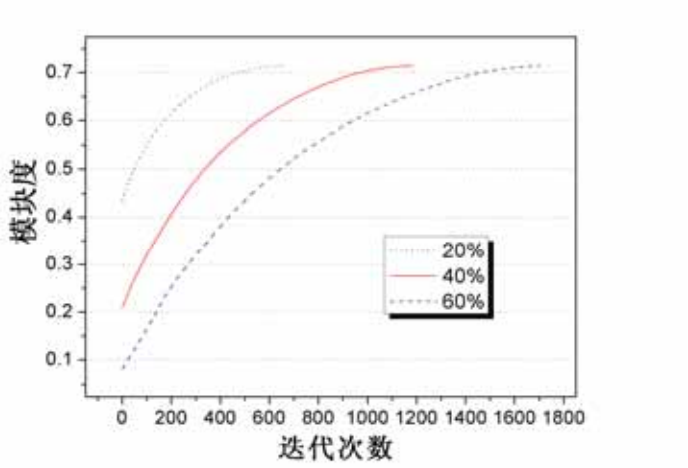


图3：对包含不同噪声比例实验数据的修正过程中模块度的变化曲线

从图3中看出，对于各种噪声比例的实验数据，通过算法的修正处理，都能得到稳定的结果。针对噪声比例分别为20%、40%、60%的实验数据进行修正，算法的迭代次数分别为656、1186、1703。从中看出算法的迭代次数与噪声数据的数量基本相当，也就是说，算法的收敛速度与噪声数量基本一致。从中可以看出，算法每一次迭代都能够准确找到噪声数据，并将其修正到合适的类别中，说明算法是高效的。

针对不同噪声数据的比例的实验数据，分别统计修正前后的模块度、准确率、召回率、以及Fmeasure，统计结果如表2所示。

表2：对不同噪声比例的实验数据的修正结果

噪声比例(%)	修正前统计结果				修正后统计结果			
	模块度	准确率	召回率	Fmeasure	模块度	准确率	召回率	Fmeasure
5	0.5548	0.9462	0.9522	0.9490	0.7136	0.9617	0.9438	0.9504
10	0.5548	0.8906	0.9021	0.8960	0.7136	0.9624	0.9478	0.9536
15	0.4940	0.8366	0.8519	0.8433	0.7138	0.9583	0.9441	0.9496
20	0.4342	0.7827	0.8017	0.7904	0.7136	0.9593	0.9433	0.9500
25	0.3715	0.7316	0.7510	0.7393	0.7140	0.9552	0.9370	0.9435
30	***0.3081***	***0.6837***	***0.7015***	***0.6900***	***0.7135***	***0.9581***	***0.9378***	***0.9451***
35	0.2593	0.6331	0.6526	0.6393	0.7139	0.9567	0.9302	0.9389
40	0.2092	0.5835	0.6016	0.5892	0.7138	0.9520	0.9320	0.9389
45	0.1826	0.5347	0.5517	0.5392	0.7139	0.9514	0.9286	0.9368
50	0.1356	0.4864	0.5008	0.4879	0.7134	0.9527	0.9268	0.9346
55	0.1089	0.4389	0.4520	0.4399	0.7137	0.9501	0.9277	0.9346
60	0.0777	0.3907	0.4013	0.3902	0.7136	0.9436	0.9230	0.9306
65	0.0440	0.3424	0.3511	0.3411	0.7123	0.9369	0.9148	0.9229
70	***0.0299***	***0.2946***	***0.3020***	***0.2932***	***0.7105***	***0.9334***	***0.8968***	***0.9089***

从实验结果中可以看出，第一、在同样噪声含量的情况下，噪声修正后的结果明显好于没有修正的情况，比如第6行噪声比例为30%，修正前的准确率是0.6837，修正后的准确率是0.9581，修正前的召回率0.7015，修正后为0.9378，其他修正后的指标都明显地好于修正前；第二、针对不同噪声比例的实验数据，通过算法修正，都能够获得准确的结果；第三、在本文的实验中，对于包含高噪声比例的数据，通过算法修正，同样能够获得很高准确率的结果，比如包含70%噪声比例的数据，修正前的准确率是0.2946，修正后

是0.9334。第四、无论是对包含不同噪声比例的实验数据，还是对多次随机生成的相同噪声比例的实验数据，本文提出的算法都能够很好地进行修正并得到比较稳定的、趋于一致的结果。说明本算法不仅具有较高的有效性，还具有较高的鲁棒性。

6 结论

粗分类数据中的噪声数据对文本分类结果产生不良影响。本文针对这个问题，提出了基于模块度指标优化的噪声数据修正算法。通过改变标准语料库中文档所属的类别，构造出含噪声的粗分类数据，并使用本文提出的算法进行了噪声修正处理。从实验结果中看出，所提算法能够准确发现噪声数据，并将噪声数据修正到正确的类别中。针对不同的粗分类实验数据，通过算法修正后能够获得准确的、稳定的结果，证明了算法的有效性和鲁棒性。本算法可被用于数据挖掘和文本挖掘研究中的数据预处理工作，或作为辅助工具用于情报文献分类整理等工作。不仅如此，还适于更加广泛的分类问题，可以推广到其他领域。

本文所提算法的迭代次数与数据质量相关，算法每次迭代选择一个数据对其进行修正。下一步工作将研究通过增加一些判定条件，每次可以选择多个数据同时修正的改进算法，这样可以减少迭代次数。同时也可以结合其他修正迭代算法，进一步提高算法效率和处理结果的质量。

参考文献

1. 陈涛，谢阳群．文本分类中的特征降维方法综述．情报学报，2005，24(6)：690-695
2. Qiong Chen. Feature Selection for the Topic-Based Mixture Model in Factored Classification. Computation Intelligence and Security, 2006 International Conference, Nov. 2006(1): 39-44.
3. Rong-Lu Li, Yun-Fa Hu. Noise Reduction to Text Categorization Based on Density fro KNN. Machine Learning and Cybernetics, 2003 International Conference. Nov. 2003(5): 3119-3124.
4. Shui-Geng Zhou, Tok Wang Ling, Ji-Hong Guan, Jiang-Tao Hu, Aoying Zhou. Fast Text Classification: A Training-Corpus Pruning Based Approach. Database Systems for Advanced Applications, 2003, Proceedings Eighth International Conference. Mar. 2003: 127-136.
5. David A. Bell, J.W. Guan, Yaxin Bi. On Combining Classifier Mass Functions for Text Categorization. Knowledge and Data Engineering, IEEE Transactions, Oct. 2005(17): 1307-1319.
6. 杨建良，王永成．基于 KNN 与自动检索的迭代近邻法在自动分类中的应用．情报学报，2004，23(2)，137-141
7. http://www.nlp.org.cn/docs/download.php?doc_id=281
8. R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, 1999.
9. 程泽凯，林士敏．文本分类器准确性评估方法．情报学报，2004，23(5)：631-636
10. M. Girvan and M.E.J.Newman. Community structure in social and biologican networks. Proc, Natl. Acad. Sci. USA 99, 7821-7826(2002).
11. M.E.J. Newman. Detecting community structure in networks. Eur. Phys. J. B 38, 321-330(2004)
12. M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical Review E 69, 026113(2004)