第41卷 第6A期
2014年6月

计 算 机 科 学
Computer Science

Vol. 41 No. 6A
June 2014

加权抽样对相似性学习算法的改进效果研究

刘欣悦 刘广钟

(上海海事大学信息工程学院 上海 201306)

摘 要 当今诸多聚类算法需要通过计算样本间距离来得到样本相似性。因此对这类算法而言,距离的计算方法尤为重要。对部分现有距离度量学习或相似性学习算法进行研究后可以发现,多数算法在选择学习样本的过程中,都采用了重复随机抽样的方式。这一抽样方式使所有训练节点都有均等概率用于度量或相似性学习,但因样本位置不同,对分类算法而言样本的分类难度也不同。如果能针对较难分类的样本进行着重学习,并适当减少对易分类点的学习时间,便能提高学习过程的效率性,减少学习过程的时间。节约时间成本,在大数据时代有不容忽视的意义。

关键词 相似性度量,距离度量,加权抽样,机器学习,k-NN,Boosting

中图法分类号 TP181 **文献标识码** A

Study on Similarity Learning with Weighted Sampling

LIU Xin-yue LIU Guang-zhong

(Information and Engineering Faculty, Shanghai Maritime University, Shanghai 201306, China)

Abstract A lot of classification algorithms get the similarity between samples according to their distance. Therefore, for this kind of algorithms, the way for getting distance is very important. Studies in existing metric or similarity learning algorithms find that most of the existing methods take use of random samples from training database for learning. This sampling method gives an equal probability for every training samples to be used for metric learning. However, the different location results in different classification difficulty of samples. If those samples who are difficult classified could be used more frequently in learning, while other samples arranged less learning time, the efficiency of learning will be improved. Reducing learning time is significant in Big-Data era.

Keywords Similarity measurement, Distance metric, Weighted sampling, Machine learning, k-NN, Boosting

分类算法是机器学习领域的重要组成部分。不同的分类方法都具有相同或相似的目的,即将特征相似的样本归于同一类,同时将特征不同的样本归于不同类别。因此,对于所有分类算法来说,辨别两个样本是否相似的方法具有决定性作用。

当前存在许多基于距离的分类方法,如 k 阶近邻算法,k-means 算法等等。两个样本间的距离衡量则可以说是这些算法的基础。众所周知,测量两个样本之间距离最简单的方法,是在欧几里得空间里获取欧氏距离。但在某些情形下我们会发现,样本之间的相似性关系,同在欧几里得空间反映出的距离并不符合。也就是说,有时相似的样本会比不太相似的样本在欧几里得空间的距离还要大,这样一来就会影响基于距离的分类算法的准确度。这种情况在两个类别边界的位置尤为容易发生。为了保证分类算法有较高的准确度,我们希望在距离度量空间中,同类的样本可以彼此较接近,而不属于同类的样本彼此距离可以远得多。这也就是距离度量学习的目标。

我们所要讨论的相似性学习与距离度量学习是非常相关的。在相似性学习时,我们希望通过一个相似性函数对两个样本进行相似性计算。对于更相关的样本,该函数能返回更大的分数,而对于更不相关的样本,则得到更小的分数。也就是说,相似性和距离值是个相对的概念:距离大的样本拥有更小的相似性、距离小的样本间有更大的相似性。所以相似性学习和距离度量学习可以看作是同一个问题。

用数学方程来表达这些概念,则如式(1)所示,其中 $S(x,y)$,$D(x,y)$ 和 $R(x,y)$ 分别代表样本之间的相似性、距离以及相关性。

$$\begin{aligned} &S(x_i,x_j)<S(x_i,x_k)\\ &D(x_i,x_j)>D(x_i,x_k)\\ &R(x_i,x_j)<R(x_i,x_k) \end{aligned} \tag{1}$$

我们知道,在不同的数据集中,数据会有不同的特征。因此我们希望样本距离的测算能够适应各个数据集的结构特征,以达到优化分类算法的目的。这也是距离度量需要学习的原因之一。

度量空间是一个集合,在其中可以定义在这个集合的元素之间的距离(叫做度量)的概念[15]。其中距离是定义在向

本文受国家自然科学基金项目(61202370),上海市教委科研重点创新项目(12ZZ151),上海市浦江人才计划项目(11PJ1404300),上海海事大学2013年研究生学术新人培育计划(工学)(GK2013077)资助。

刘欣悦(1989—),女,硕士生,主要研究方向为机器学习,E-mail:xinyue-50@163.com;刘广钟(1962—),男,博士,教授,博士生导师,主要研究方向为分布式数据库、分布式人工智能、计算机网络技术、网格计算、CIMS 技术、物流信息化技术等。

量空间中的一种函数，指(两物体)在空间或时间上相隔或相隔的长度[16]。也就是说，我们可以将某个数据集合中的所有对象，投影到一个新的度量空间中。在这一度量空间中，同类的或相似的物体之间距离很小，而不同类的或不相似的物体之间距离较大。式(2)是马氏距离的表达法，其中有一个矩阵 M，它就是数学概念上我们需要学习的度量空间。这一度量考虑到了数据集的特征和各个样本之间的关系，能使经由它计算出来的距离值符合我们所期望的要求。在理想情况下，同类点之间的距离比不同类点之间的距离小得多，具体可以参见图1。

$$D_M(x,y)=\|x-y\|_M^2=(x-y)^{\mathrm{T}}M(x-y) \tag{2}$$

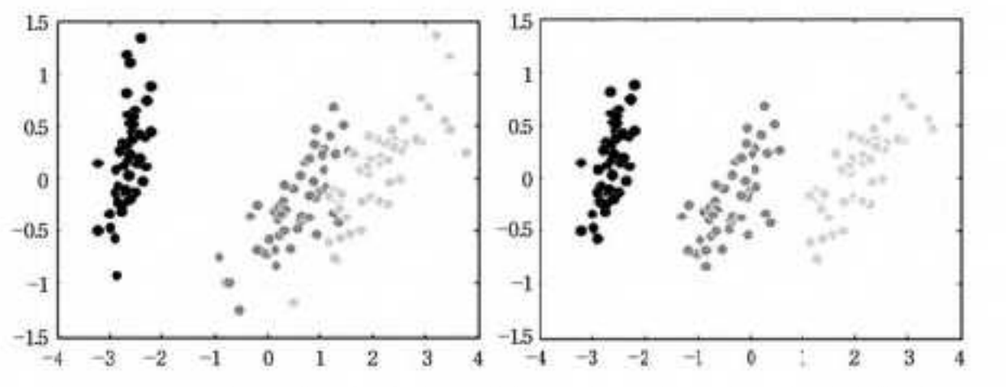

图1 Iris数据库在训练前和期望训练后的二维空间内投影

在图1中，我们使用了四维的Iris数据库，为了使其在二维空间中呈现，使用PCA将其投影。需要说明的是，在马氏距离(式(2))中，矩阵 M 通常是一个半正定矩阵，当数据集的维度为 d 时，M 则是一个 $d\times d$ 矩阵，而它可以被分解，其形式如等式(3)所示，其中实数矩阵 L 则可以用来对数据进行投影。

$$M=L^{\mathrm{T}}L \tag{3}$$

综上所述，可以知道在相似性学习或距离度量学习中，矩阵 M 是我们需要学习的结果，它应该是一个适应数据库特征，并且能够提高分类算法准确率的度量空间矩阵。

1 现有研究

当前在相似性学习或距离度量学习领域已经存在许多方法，其中大多数都是基于马氏距离的。许多方法都在试图学习接近理想化的矩阵 M。因为 M 是个半正定矩阵，所以 M 的优化过程可以被当作一个凸规划问题求解。然而，要保证矩阵 M 是一个半正定矩阵，相当于在运算方面增加了复杂度，因此许多方法也采用了直接学习 L 的方式，L 如式(3)中所示。无论具体采取哪一种方法，其过程一般都是先设定一个学习目标，根据这一目标设计以训练数据和目标学习矩阵为参数的目标方程和约束条件，并用数学手段对目标学习矩阵进行优化。以下将会简要介绍几种现有的相似性或距离空间学习方法。

1.1 OASIS

OASIS是可扩展的图片相似性在线学习法[1]。其学习目标是获得一个相似性函数 $S_W(p_i,p_j)\equiv p_i^{\mathrm{T}}Wp_j$，使其返回值与输入的两幅图片的相关度成正相关。其中 W 即为所需要学习的目标相似性度量矩阵。该方法特点是采取了在线学习方法，每次学习只采用一个三元组 (p_i,p_i^+,p_i^-) 作为输入。该三元组中 p_i 是某个查询图片，p_i^+ 是与 p_i 相关的或属于同类的图片，而 p_i^- 是与 p_i 不相关的或不属于同类的图片。

采用在线学习的方法优势为：(1)当有新的数据加入训练集时能够以较低成本适应新训练数据；(2)即使只用一部分训练集内的数据也可以获得比较优化的训练结果。同时，用三元组训练的好处是，训练过程不需要知道训练集的具体分类信息，只需要知道三个元素之间的相对相关性即可。另外，因该方法强调相似性学习而非距离度量学习，所学的目标矩阵 W 并不要求一定对称，也不要求一定是正值，这样一来大大减少了优化时的计算量。

但是这一算法也有一些不足。因其是针对图片相似性的学习算法，在对数据的形态上，它假设了所输入的图片信息数据是一个稀疏矩阵。在符合这种条件的情况下，其学习速度较快。但当输入数据不是这种结构时，其学习速度受到影响。所以它对输入的训练集有一定结构要求。

1.2 LMNN

大间隔近邻算法[2]，其可以看作是OASIS算法的一个重要思想基础。它尝试拉大不同类别点的间距，并且同时缩短同类点之间间隔。它主要用于提高K阶近邻算法的正确率，因此基于kNN的特点提出了两个约束点：首先每个训练数据点都要和它最近的 k 个数据点有同样的类别标签；其次不同类别标签的数据点之间都有大间距隔开。

根据这两个约束，这一方法的损失函数分两部分，分别用于惩罚同类间的过大距离和不同类间的过小距离。目标函数则旨在最小化全局损失函数值。损失函数中，用于惩罚不同类间过小距离的部分与OASIS的损失函数十分类似。

LMNN算法采用的不是在线学习法，因此其学习时需要同时使用整个训练集的数据作为输入，但经过修改，也可以获得其在线学习的方法。另外，它的学习过程中用到了分类标签，因此提供的训练集信息要求相对较高，不只是具有彼此之间的相对关联程度即可，还需要每个数据点的具体分类信息。

1.3 ITML

信息理论度量学习方法[3]是运用信息理论，将距离度量学习问题转换为最小化两个多元高斯模型的微分相对熵的问题。其利用的信息理论为：在一个马氏距离空间和一个等均值多元高斯分布领域存在一个简单的双射。用这一理论，它就把一个距离表达式转换成一个有马氏距离中距离空间矩阵的高斯分布表达式。在优化过程中，最小化被训练空间对应的分布和初始空间对应的分布的Kullback-Leibler差异即可。最小化这一差异是为了使在训练过程中训练目标矩阵的变化能比较平缓。大多数方法，包括OASIS和LMNN都考虑到这一方面。

这一方法的特点是能够适应不同的约束条件。它的目标函数如上所述，比较容易适应不同的约束，因此使得这个方法比较灵活。其不要求特征值的计算以及半定编程过程，使其成为一种相对而言有效率的学习方法。但鉴于其优化更新过程的算法复杂度为 $O(cd^2)$(其中 c 是约束的数量，d 是训练集的维度)，因此当训练集维度很大时，它的算法复杂度会大大上升。所以这种方法在训练集维度较小时比较推荐。

1.4 其他方法

除以上3种方法外，关于距离度量和相似性的学习还有许多种算法。

其中有些也采用类似的理论，但是应用有所不同。比如统一距离度量学习(UDML)方法[4]，其主要解决图像标签问题。它同时利用了图像的视觉信息和文本信息，使二者结合。其中视觉信息的训练部分同LMNN十分类似，而文本信息训

练部分则采用了余弦相关性理论。二者结合为目标函数,同时训练图片的两种信息。这种方法就是针对具体应用的。

有些则采用了十分不同的想法。例如挤压类度量学习(MCML)方法[5]。它依赖于一个空间理想化构想,即把所有同类别的数据都压到一个点上,不同类别的点则相隔很远。其优化也使用了 Kullback-Leibler 差异。在训练过程中,如果训练集新增加了一个类别或者一个数据,该方法就需要计算新增数据同其他每个数据之间的距离,计算量很大。有一种与 MCML 比较相似的模型 NCM[6],该方法在面对新增数据时,只需要计算新增数据同每一个类别中心的距离。因此,若训练集有被更新的可能,选择 NCM 方法比较有效率。

2 加权抽样方法

以上所有我们所介绍的方法中,有在线学习的方法,也有非在线方法。无论哪种学习方法,其学习的过程都涉及到训练数据抽样的问题。对于训练结果而言,用于训练的数据无疑是重要的。而现存方法中,大多数方法(包括 OASIS,ITML,MCML…)采用的抽样方式都是随机抽样。这就意味着,在训练集中的所有数据项,都有相同的几率被选中进行训练。但是从图 1 的左图中可以看出,在一个训练集中,有些点的位置在自己分类的主体区域,且离其他分类较远,则比较容易被分类算法正确分类。但有一些点的所在位置使其难以被分类,尤其是在类的边界,或者两个类交界处的点。我们学习新的度量空间,就是为了在这一新的度量空间中,能减少容易令基于距离的分类算法混淆的情况。因此我们希望能够针对那些容易混淆的数据点进行更多的训练,而比较容易分类的数据点则花更少的精力来训练。我们认为,按照这种想法进行的训练能够大大提高训练算法的效率。

为了把数据点根据分类难度安排不同的训练频度,我们采用了加权的方式,而分配权重的方法则借鉴了 Boosting 算法。Boosting 算法简单说来,就是提高产生错误的样本的权重,并且降低未产生错误的样本的权重。抽样时根据权重进行选择,权重高的样本更容易被抽中,权重低的样本更不容易被抽中。而样本是否产生错误,则根据各个算法中的损失函数可以进行判断。

伪代码算法 1 中,我们以在 OASIS 算法中增加加权抽样部分为例,展示了如何在程序中实现加权抽样。初始化时,我们对每个三元组给予一个标签为 0,并且给予相同的权重。在学习过程中,将获得损失值大于 0 的三元组的标签改为 1,而获得损失值为 0 的三元组则直接将权重减为 0。在 m 次(m 为一个常数,实验中我们使用的是 20,并且得到了不错的结果)的学习后,针对之前标记过 1 的三元组进行加权。当算法满足 OASIS 的终止条件,或者权重总和为 0 时,算法终止。

算法 1 Pseudo-code of Weighting sample part in OASIS

```
Initialization:
  W0 = I
  n = number of triplets
  weights[n] = {1/n, 1/n, …, 1/n}
  records[n] = {0,0,0,…,0};
Iteration:
  Repeat
    At ith iteration, samples a triplet whose index is j;
    Get loss value;
If loss>0 then
    records[j] = 1;
else
    weights[j] = 0;
end if
if i mod m == 0 then
err = weights * records;
    err = min (0.5, err);
    c = log((1-err)/err);
    c = (1/2) * c;
    weights = weights. * (exp (c * records))';
    weights = weights/Σ(weights);
    records[n] = {0,0,…,0};
  end if
until Convergence or sum(weights) == 0;
```

我们相信,通过这种方式进行抽样,能大大提高训练效率,使训练过程尽快收敛。之后我们将通过实验加以证实。

3 实验及结果

3.1 评价方法

在修改了程序以后,为验证我们的假设是否正确,进行了一些相关实验。实验对 OASIS 算法和 ITML 算法加以改进,使其抽样方法由随机抽样改为加权抽样。为了对比修改前后的效果,我们在修改前后训练获得的度量空间中进行分类测试,使用的分类方法为 K 阶近邻法,其中 K 为 3。为了获得更为真实的实验结果,在训练和测试时,我们采用了交叉验证(Cross-Validation)的方式,把测试数据库分为 5 个文件夹进行。每一种方法下的每一个数据库,进行 50 次测试,最后以平均值作为实验结果。

除了正确率以外,还搜集了运行时间、学习迭代次数作为实验比较数据。因为需要验证的主要是在学习效率上的提高,因此时间和迭代次数是必不可少的数据。

3.2 数据库

在实验中使用了 10 个数据库,它们都来自 UCI 机器学习存储库的分类算法数据库中。这 10 个数据库分别是:Iris,Wine,Soybean-large,Balance-Scale,Ionosphere,Seeds,Blood-Transfusion,Australian-Credit-Approval,Car 以及 Breast-cancer。它们所含有的数据量和属性维度如表 1 所列。

表 1 实验所用 10 个数据库的数据数量和属性数量

	Ir~	Wi~	So~	Ba~	Io~
数据数量	150	178	307	625	351
属性数量	4	13	35	4	34
	Se~	Car	Au~	Tr~	Br~
数据数量	210	1728	690	748	699
属性数量	7	6	14	5	10

3.3 实验结果

首先,为了验证算法是否能够针对边界等比较难以分类的点提高权重,我们用 Iris 数据库和 OASIS 算法做了一次训练。训练过后,我们把每个三元组的权重分配到里面的三个数据项上,每个数据项的权重则为其所在所有三元组权重之和。分配权重后,我们将 Iris 数据库里的点再次在二维空间中绘出,且每个点的大小与自身的权重大小成正比。这样一来,直观地通过数据点的大小就可以看出它在经过训练以后

权重的大小。绘图结果见图 2。

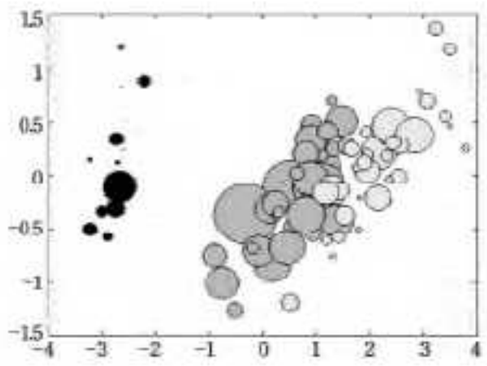

图 2　训练后投影到二维空间的 Iris 数据，数据点大小与其训练后的权重成正比

由图 2 可以观察出，最左边的类别由于离另外两个分类较远，大多属于比较容易区分的数据，所以其训练后的权重普遍偏小。而靠右的两个类别彼此靠近，尤其是分界线周围的点权重都要大得多。因此可以说，我们的加权方法对边界周围难以分类的点，能够起到重点训练的作用。

接着就可以进行具体实验数据的比较。如前两节所述，我们先用 OASIS(O) 和 ITML(I) 原本的随机抽样的方法，对 10 个数据库分别进行交叉验证，获得 50 次实验后的平均值。再将两种方法改为使用加权抽样的方式，获得数据，进而进行比较。实验结果列于表 2 中。

表 2　OASIS(O) 和 ITML(I) 在使用加权抽样前后的实验结果对比

数据库	学习算法	Random			Weighted			Difference		
		acc	iter	time	acc	iter	time	acc	iter	time
Ir	O	96.70	10000	27.70	97.04	511	1.79	0.34	−9489	−25.91
	I	96.52	29715	7.26	96.63	3363	2.96	0.11	−26352	−7.15
Wi	O	94.38	10000	23.78	96.76	549	2.36	2.38	−9451	−21.42
	I	97.88	29170	8.13	97.74	6792	5.12	−0.14	−22378	−3.01
So	O	89.12	10000	54.11	91.47	10000	65.21	2.35	0	11.1
	I	91.51	99429	46.44	90.58	91443	74.32	−0.93	−7986	27.88
Ba	O	91.25	10000	17.81	91.69	451	2.16	0.44	−9549	−15.65
	I	91.31	51296	12.36	90.89	8061	5.84	−0.42	−43235	−6.52
Io	O	85.24	10000	22.85	84.22	179	0.70	−1.02	−9821	−22.15
	I	86.02	944	0.55	86.21	947	0.85	0.19	3	0.3
Se	O	90.51	10000	14.82	90.21	454	1.53	−0.3	−9546	−13.29
	I	90.71	44589	10.69	91.18	5363	3.54	0.47	−39226	−7.15
Tr	O	73.66	6794	12.26	75.26	8000	31.84	1.6	1206	19.58
	I	74.53	553	0.25	74.27	200	0.24	−0.26	−353	−0.01
Au	O	68.04	6646	12.10	70.46	189	0.84	2.42	−6457	−11.26
	I	66.60	1535	0.49	67.04	375	0.34	0.44	−1160	−0.15
Ca	O	91.15	10000	21.49	86.35	1126	7.02	−4.8	−8874	−14.47
	I	92.44	92570	22.33	92.44	22963	15.54	0	−69607	−6.79
Br	O	94.57	10000	25.23	94.75	157	0.82	0	−9843	−24.41
	I	96.25	2290	0.65	96.27	734	0.63	0.02	−1556	−0.02

表 2 记录的是 OASIS 和 ITML 算法对 10 个数据库的实验结果，其中分别记录了它们使用随机抽样和加权抽样时的正确率、迭代次数、运行时间，并比较两种抽样下的实验数据区别。在差别部分，粗体标明加权抽样得到了更好的结果，非粗体标明加权抽样后更不好的结果，0 则表示没有变化。

从中我们可以得出一些统计分析，具体有以下两点：

1)以上所有实验数据中，平均正确率的变化值为 0.1445。这一正值说明，当抽样算法由随机改为加权以后，对分类算法的准确率影响总体没有下降，甚至还有少量提高。另外可以看出，使用加权抽样方法后，65% 的正确率等于或高于使用随机抽样方法的正确率。

2)平均迭代次数和运行时间上的差别值分别为 −14183.7 和 −6.025。两个负值说明，无论在迭代次数上还是运行时间上，使用加权抽样都比使用随机抽样有所下降。另外，80% 的结果中运行时间和迭代次数均有减少。

因此，我们可以说，加权抽样方法在一般无正确率损失的前提下，减少了 OASIS 和 ITML 算法的运行时间和迭代次数。也就是说，使用加权抽样方法，能够使这两个算法的学习效率大大提高。

结束语　在前人提出了种种相似性学习和距离度量学习算法的环境下，我们经过思考研究，尝试在算法上进行一些改进。我们把关注点放在训练时的训练样本抽样方法上，将以往常用的随机抽样训练改成加权抽样训练后，利用 OASIS 和 ITML 算法进行了一系列的实验。实验结果表明，加权抽样能给这两种算法带来学习效率上的提高，缩短学习时间并减少学习迭代次数，同时不产生或产生极少量的分类正确率的损失。

但我们还有许多工作需要继续学习。首先，我们仅尝试了两个度量学习算法上的改进，将来的学习中，可以进一步在其它方法上进行实验。其次，在训练中，我们将没有产生损失值的样本的权重直接改为了 0，今后的实验中，我们可以尝试对这类样本逐渐减少权重的方式，并进一步实验，查看结果是否会不同。最后，对于加权算法中 m 的选择，我们可以尝试使其根据数据库的规模而变化，即设计一个以数据库规模为输入，m 为输出的函数，使改变样本权重的行为更灵活地适应不同的数据库。

参 考 文 献

[1] Chechik G, Sharma V, Shalit U, et al. Large scale online learning of image similarity through ranking[J]. The Journal of Machine Learning Research, 2010, 11: 1109-1135

[2] Blitzer J, Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification[C]// Advances in Neural Information Processing Systems. 2005: 1473-1480

[3] Davis J V, Kulis B, Jain P, et al. Information-theoretic metric learning[C]// Proceedings of the 24th international conference on Machine learning. ACM, 2007: 209-216

[4] Wu P, Hoi S C H, Zhao P, et al. Mining social images with distance metric learning for automated image tagging[C]// Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. ACM, 2011: 197-206

[5] Globerson A, Roweis S T. Metric learning by collapsing classes [C]// Advances in Neural Information Processing Systems. 2005: 451-458

[6] Mensink T, Verbeek J, Perronnin F, et al. Large scale metric learning for distance-based image classification[R]. 2012

[7] Kulis B, Sustik M, Dhillon I. Learning low-rank kernel matrices [C]// Proceedings of the 23rd international conference on Machine learning. ACM, 2006: 505-512

[8] Demšar J. Statistical comparisons of classifiers over multiple data sets[J]. The Journal of Machine Learning Research, 2006, 7: 1-30

[9] Liu Yang, Rong Jin. Distance metric learning: A comprehensive survey[D]. Michigan State Universiy, 2006: 1-51

[10] 张丽娟，李舟军. 分类方法的新发展：研究综述[J]. 计算机科学，2006，33(10)：11-15

（下转第 405 页）

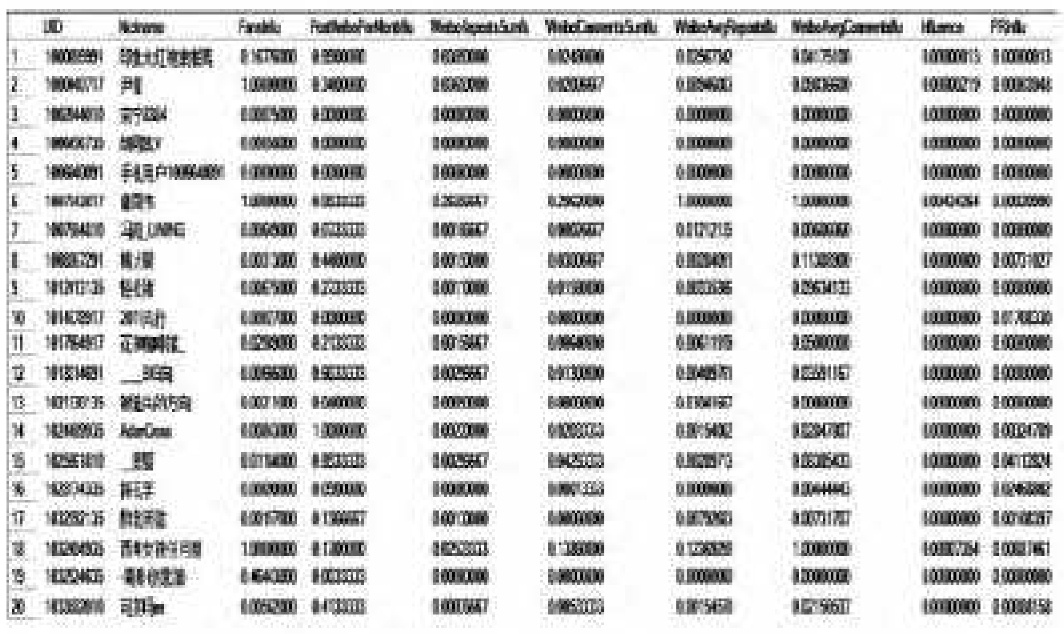

图 4 贝叶斯-PageRank 实验结果

图 4 为运用贝叶斯-PageRank 算法对实验集合数据进行处理后，用户影响力排名的最终结果。从结果可以明显地看出，将该算法运用于实验集合节点后得到用户节点“印象大红袍衷柏夷”、“尹雄”、“宋宁 8384”、“胡明凯 V”、“曹国伟”等排名较高。

对排名靠前的用户节点，我们逐一核实了其微博网站的属性及粉丝的状况。结果发现，重要节点用户集合的选取结果较为符合微博网络实际情况，另一方面重要节点之间的相对次序同样具有一定的准确率。故此算法针对微博社会网络中重要用户节点的筛选切实有效。

5 引导网络舆情

微博社会网络发展至今，给人们生活带来便捷的同时，也对网络安全造成了威胁。微博社会网络中的重要用户节点对维护网络安全至关重要。本文对微博社会网络重要用户节点进行筛选，通过分析重要用户节点，对网络舆情引导提出如下 3 个对策：

. 微博社会网络中的重要用户教育

对上述研究得到的微博社会网络中的重要用户进行持续教育，树立其维护网络安全的意识，加强其作为重要用户节点的社会责任感。开展网络道德教育的同时，加强网络法制教育。教育重要用户不散步谣言和反动言论，不诽谤他人，不传播有害信息和病毒，维护网络资源建设的安全。

. 加强舆情信息传播应急处理能力

舆情信息的传播速度快、传播范围广、影响力大，加强舆情信息传播应急处理能力，是为了避免失真舆情信息的传播对国民经济及生活造成危害。一方面要识别舆情信息中的虚假信息，及时辟谣，维护社会网络稳定；另一方面要正确面对舆情信息反映的问题，公正公平公开进行处理。

结束语 本文首先设计并实现了新浪微博专用爬虫，获得研究数据。其次通过实验得到重要用户节点指标，提出贝叶斯-PageRank 算法并利用其筛选重要用户节点，实验验证了重要用户节点的有效性。最后，通过对重要用户节点的监测实现网络舆情发现并给出相关舆情引导策略。

参 考 文 献

[1] Kwak H，Lee C，Park H，et al. What is Twitter，a social network or a news media[C]// Proceedings of the 19th International Conference on World Wide Web. 2010：591-600

[2] Yang Jiang，Counts S. Predicting the Speed，Scale，and Range of Information Diffusion in Twitter[C]// Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. 2010：355-358

[3] Cha M，Benevenutob F，Ahn Y-Y，et al. Delayed information cascades in Flickr：Measurement，analysis，and modeling[J]. Computer Networks，2012，56(3)：1066-1076

[4] Tang Si-yu，Blenn N，Doerr C，et al. Digging in the Digg Social News Website[J]. IEEE Transactions on Multimedia，2011，13(5)：1163-1175

[5] Xu Zhi-heng，Zhang Yang，Wu Yao. Modeling User Posting Behavior on Social Media[C]// Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2012：545-554

[6] Leavitt A，Burchard E，Fisher D，et al. The influentials：New approaches for analyzing influence on twitter[R]. Web Ecology Project，2009

[7] Brown P，Feng Jun-lan. Measuring User Influence on Twitter Using Modified K-Shell Decomposition[C]// Fifth International AAAI Conference on Weblogs and Social Media Workshop on the Social Mobile Web Brown. 2011：18-23

[8] Wang Jian-wei，Rong Li-li，Guo Tian-zhu. A new measure of node importance in complex networks with tunable parameters[C]// 4th International Conference on Wireless Communications，Networking and Mobile Computing，2008(WiCOM' 08). 2008：1-4

[9] Heidemann J，Klier M，Probst F. Identifying Key Users in Online Social Networks：A PageRank Based Approach[C]// ICIS 2010 Proceedings. 2010：79-100

[10] Zhang Meng，Sun Cai-hong，Liu Wen-hui. Identifying Influential Users of Micro-blog Services：A Dynamic Action-Based Network Approach[C]// PACIS 2011 Proceedings. 2011：223-236

[11] Hallberg V，Hjalmarsson A，Puigcerver J. TweetRank：An adaptation of the PageRank algorithm to Twitter world[C]// Search Engines and Information Retrieval. 2012：1-11

[12] 刘华. 网页信息抽取及建库系统 C# 实现[J]. 计算机工程，2006，32(16)：213-216

[13] 王一姝，郭海峰. 网络 IP 流量的自相似性分析与研究[J]. 科技信息，2013(3)：127

(上接第 390 页)

[11] He X，King O，Ma W Y，et al. Learning a semantic space from user's relevance feedback for image retrieval[J]. Circuits and Systems for Video Technology，IEEE Transactions on，2003，13(1)：39-48

[12] Peng J，Heisterkamp D R，Dai H K. Adaptive kernel metric nearest neighbor classification[C]// 16th International Conference on Pattern Recognition. IEEE，2002，3：33-36

[13] Bar-Hillel A，Hertz T，Shental N，et al. Learning distance functions using equivalence relations[C]// Proc. International Conference on Machine Learning，2003

[14] Bar-Hillel A，Hertz T，Shental N，et al. Learning distance functions using equivalence relations[C]// ICML. 2003，3：11-18

[15] http://zh.wikipedia.org/wiki/%E5%BA%A6%E9%87%8F%E7%A9%BA%E9%97%B4 [OL]

[16] http://zh.wikipedia.org/wiki/%E8%B7%9D%E7%A6%BB [OL]

[17] Duda R O，Hart P E，Stock D G. Pattern Classification (Second Edition)[M]. 北京：机械工业出版社，2010：143-155