基于 SOM 聚类的可视化方法及应用研究

刘 芳
(大连交通大学 软件学院, 辽宁 大连 116028)

摘 要: 提出了用无监督的自组织映射方法对金融数据进行聚类,并用平行坐标和交互式的圆形平行坐标方法在二维平面上表示出来。用这种方法形成清晰的可视化聚类结果,不仅有效地总结了数据特征,还提高了聚类的可视效果,从而便于发现数据的变化趋势。

关键词: 聚类; 平行坐标; 金融数据; 可视化分析; 自组织映射
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)04-1300-04
doi:10.3969/j.issn.1001-3695.2012.04.027

Research on visualization method and its application based on SOM clustering

LIU Fang, TIAN Kai, ZHOU Zhi-guang, LIN Hai
(*Software Institute*, *Dalian Jiaotong University*, *Dalian Liaoning* 116028, *China*)

Abstract: This paper proposed an unsupervised SOM-based cluster method for dealing with financial data. The clustered result was visualized on two dimensions by using parallel coordinate and interactive circular parallel coordinate. With this method, it can form a clear visualization of the clustered result, improve the effect of visual clustering and enable user to reveal underlying patterns.

Key words: clustering; parallel coordinate; financial data; visual analysis; SOM(self-organizing map)

0 引言

如今,多维数据普遍存在,特别是在经济领域中,每天的业务都会产生大量的数据。虽然目前数据库系统可以有效地实现数据的录入、查询、统计等功能,但很难辅助金融分析师发现数据中存在的关系和规则以及预测未来的发展趋势。因此,分析和理解复杂多维金融数据集是一项困难而有挑战性的工作。

在金融数据可视化中,比较传统的可视化方法有柱状图、饼图、曲线图等[1,2]。其中最著名的是用 K 线图来反映股票价格变化趋势。这些传统的可视化方法不适应于海量的金融数据。近几年,一些新的可视化技术被提出并应用于金融领域。TreeMap 可视化技术在金融领域广泛应用,如市场系统图[3]。在 TreeMap 视图中用区域大小表示市场份额,用颜色编码表示资产回报率。Theme River[4] 也被应用于金融时间序列数据,以确定资产价格的增加或减少。基于小块的平行坐标的方法提出了在绘图区域划分成长方形小块[5],每一个小块表示在这个小块中与该小块交叉的折线的总数,即为一个交叉密度。该交叉密度又通过传输函数映射颜色和透明度。该方法已应用于现实世界的数据集——2006 年美国股票共同基金数据,为投资共同基金经理、金融规划者和投资者提供了一个投资分析工具。文献[6~8]用轨迹图来表示股票指标的变化轨迹,在一定时期内,资产回报的分布通过增长矩阵的方法进行可视化并分析[9]。然而,上述可视化方法比较复杂,不宜于辅助用户迅速地发现投资机会,尤其针对海量的金融数据,具有很大的局限性。因此本文提出了平行坐标[10]及圆形平行坐标[11]的可视化方法。平行坐标能将多维数据映射到一组平行的等距离坐标轴上,同时显示海量的多维金融数据,帮助用户快速地找到发展趋势;圆形平行坐标是对平行坐标的改进,有效地显示了数据的趋势,有利于帮助金融领域用户快速地分析初始的海量金融数据。

为了表示数据间的关系,提取隐藏的、未知的或验证已知的规律,聚类分析等成熟的方法被引入到金融数据分析领域。聚类是指按照事物的某些属性,把事物聚集成类,使类之间的相似性尽量小,类内的相似性尽量大。针对金融数据的聚类方法有很多,如 K-means、模糊聚类、BP 网络、支持向量机、SOM 等。

K-means 聚类算法[12]又称为 K-均值聚类算法,其优点是原理简单、算法速度快、伸缩性好。K-means 算法是典型的基于距离的聚类算法,采用距离作为相似性的评价指标,即认为两个对象的距离越近,其相似度就越大。该算法认为簇是由距离靠近的对象组成的,因此把得到紧凑且独立的簇作为最终目标。Das[13]已把 K-means 算法应用于对冲基金的分类中。

模糊聚类[14]是从模糊集的观点来探讨事物数量分类的一类方法。模糊聚类分析是以传统的聚类分析为理论基础,按待辨识对象属性的亲疏关系进行软划分的一种多元统计方法。它把一个没有类别标记的样本集按某种准则划分成若干个子集(类),使相似的样本尽可能归为一类,而不相似的样本尽量划分到不同的类中。该理论试图对主观的推测进行分类,如人们所说的“好”“不好”和“非常好”,并且指明得出结论的不同程度的可能性。因此,模糊聚类比精确聚类能提供更多的关于数据结构的信息。刘晓勇等人[15]利用模糊聚类方法,根据 2004 年广东省部分城市金融机构的基本财务——本、外币的存贷款情况,对各市金融机构营业状况进行探讨, 为比较、分

收稿日期: 2011-08-24; **修回日期**: 2011-10-09
作者简介: 刘芳(1976-),女,辽宁大连人,硕士,主要研究方向为可视化、数据挖掘等(maggie_liufang@126.com).

析金融机构的财务状况提供了一个有效手段。

BP 网络[16]是人工神经网络中前馈型网络的核心部分,它由一个输入层、若干隐含层和一个输出层构成。隐含层是为网络能学习到给定的样本而提供足够的可调连接权值。这些可调连接权值在学习迭代过程中,从其可能组合数空间中凑成了一组能同时满足各样本对的连接权配置方案。而后在网络工作阶段有待测试样本输入时,即按类似输入产生类似输出的相近原则,内插或外推出所需要的输出来。Zhou 等人[17]通过用 BP 网络的方法预测股票市场未来的趋势,显示出 BP 网络是一种比较好的方法。

支持向量机[18]是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的,根据有限的样本信息,在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折中,以期获得最好的推广能力。Shin 等人[19]应用支持向量机对企业破产进行了预测的实例研究。

SOM[20]网络是由 Kohonen 教授提出的对神经网络的数值模拟方法,是人工神经网络的重要分支之一。SOM 是对生物神经系统进化过程的计算机模拟,能将任意维输入模式在输出层映射成一维或二维图形,并保持其拓扑结构不变;能够根据样本出现在输入空间的概率密度,自组织地形成与这个概率分布密度相对应的神经元空间分布密度关系,是一种自组织和自学习的网络。SOM 网络模型是一种上市公司分类的新方法,并已应用到实际中。Tomas 等人[21]利用 SOM 方法,以 77 家上市公司的财务业绩为例进行了分析,得到了较好的效果。Brunno 等人[22]提出了用 SOM 进行特征聚类,并将其应用在金融投资选择上,取得了很好的结果。

以上的这些方法各有特点,在实践中也有不同的应用,但是并不能完全地适用于金融数据。SOM 由于其自组织学习的特点和聚类功能,适合金融数据的聚类,在金融数据模式识别上具有保持分类标准客观性的作用。同时,SOM 将模式上相近的样本归为一类,实际上是一种特征提取功能,能够将金融数据的结构性特征提取出来。SOM 聚类算法在金融数据分析中广泛使用[23],所以本文提出了基于 SOM 的金融数据可视化方法。

1　基于 SOM 的金融数据分类模型

尽管上述可视化方法可以有效地表示分类后的金融数据,但是仍然无法满足金融分析师的需求,给普通投资者对金融数据进行进一步的分析带来很大不便。因此,本文在采用传统的 SOM 聚类方法基础之上,利用平行坐标和圆形平行坐标的方式对分类后的数据进行可视化,以帮助金融分析师及投资者更为简单、高效地了解初始海量金融数据,为其进一步分析和投资提供了依据。

1.1　基于 SOM 聚类方法

SOM 是一种具有聚类功能的人工神经网络。它是一种无导师学习策略,学习时只需输入训练模式,网络就会对输入模式进行自组织,达到识别和分类的目的。它的基本思想是网络竞争层各神经元竞争对输入模式的响应机会,最后仅一个神经元成为竞争的胜者,并对那些与获胜神经元有关的各连接权朝着更有利于它竞争的方向调整。

SOM 描述了从一个高维输入空间到低维空间的映射。其基本结构分为输入和输出(竞争)两层。输入层神经元与输出层神经元为全互连方式,且输出层中的神经元按二维形式排列,它们中的每个神经元代表了一种输入样本。所有输入节点到所有输出节点之间都有权值连接,而且在二维平面上的输出节点相互间也可能是局部连接的。

在训练开始前,输出节点被赋予一个很小的随机权值。输入样本后,使输出节点进行竞争,并调整获胜节点及其邻域内节点的权值。训练完成时,输出层节点分布能够较好地保留数据在原空间中的拓扑分布状况。

SOM 网络的训练过程采用自组织竞争学习原则。输入为 n 维向量 $\boldsymbol{X}=[X_1,X_2,\cdots,X_N]$。SOM 算法的训练过程如下:

a)给输出层每个节点赋予初始权值,由随机函数产生。

b)输入一个新样本,即多维金融数据向量。

c)计算输入样本与每个输出节点之间的欧几里德距离,并选取一个与样本有最小距离的输出节点 i,如下:

$$d_i=\sum_{i=1}^{n}[x_j(t)-w_{ij}(t)]^2 \tag{1}$$

其中:$x_j(t)$表示输入样本在时刻 t 的值。

d)按式(2)修改输出节点 i 及其相邻神经元的连接权值,即

$$W_{ij}(t+1)=W_{ij}(t)+\eta(t)[x_i(t)-w_{ij}(t)] \tag{2}$$

其中:$i\in S_i(t)$,$1\leqslant j\leqslant n$,$\eta(t)$为学习系数,$0<\eta(t)<1$,且随时间逐渐变为零。

e)重复步骤 c)d)的学习过程。

算法中的神经元 i 是根据最小欧几里德距离来选择的。实际应用中也可以改成以最大响应输出作为选择依据。

在学习过程中,学习系数 $\eta(t)$及邻接集合 $S_i(t)$的大小是逐渐变小的。$\eta(t)$的初始值为 $\eta(0)$可以较大,随着学习时间的增长而减小,常用的计算公式如下:

$$\eta(t)=\eta(0)(1-t/T) \tag{3}$$

其中:t 表示当前迭代次数,T 为整个迭代的设定次数。邻接集合的大小也随学习过程的迭代而减小,设开始的邻域宽度为 $S_i(0)$,随迭代次数减小的计算如式(4)所示。

$$S_i(t)=S_i(0)(1-t/T) \tag{4}$$

1.2　基于 SOM 的金融数据分类

本文所选取的数据来自于国泰君安 CSMar 数据库中 2008 年 951 家工业上市公司的年报信息。分析中选取了流动比率、资产负债率、应收账款周转率、总资产周转率、净资产收益率、总资产报酬率、净利润增长率、总资产增长率[24]这八项指标,可以较全面准确地反映企业的财务状况,体现了企业的偿债、运营、盈利和成长能力。

首先对这些数据进行处理,利用 SOM 对公司进行聚类分析[25],将公司进行分类。经过归一化处理后的数据,将其输入网络进行训练,经过计算调整,最终取分类数为 6,并为每类设置了相应的颜色(见电子版),如图 1 所示。聚类颜色的选择是经过精心设计的,颜色的对比性好,有利于用户区分各类,增加各类在视觉上的差异性。

对每一类企业的八个指标求平均值,并求出每个指标总体的平均值。用户可以根据投资侧重方向不同,对各个指标设置权重值。根据《国有资本金效绩评价规则》及财务比率综合分析法[26],给不同的变量赋予不同的权重:偿债能力占的权重为

22%（其中流动比率权重为10%，资产负债率权重为12%）；营运能力占的权重为18%（其中应收账款周转率权重为9%，总资产周转率权重为9%）；盈利能力占的权重为42%（其中资产报酬率权重为12%，净资产收益率权重为30%）；成长能力占的权重为18%（其中净利润增长率权重为9%，总资产增长率权重为9%）。表1是各类别的指标平均值。

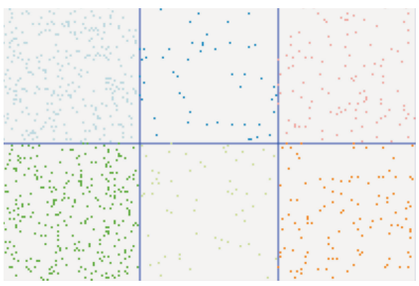


图1 SOM聚类图

表1 各个类指标平均值

类别	流动比率	资产负债率	应收账款周转率	总资产周转率	资产报酬率	净资产收益率	净利润增长率	总资产增长率
权重/%	10	12	9	9	12	30	9	9
1	1.596	0.536	54.765	0.770	0.041	-0.054	-1.577	0.096
2	1.750	0.517	31.690	0.780	0.063	0.043	-1.777	0.121
3	1.473	0.528	217.123	0.889	0.056	0.057	-0.591	0.144
4	1.676	0.526	35.322	0.869	0.050	0.066	-1.232	0.089
5	1.587	0.529	20.545	0.903	0.047	0.018	-2.336	0.050
6	1.833	1.471	952.211	0.797	0.053	0.017	-1.403	0.233
总体平均值	1.653	0.685	218.609	0.835	0.051	0.024	-1.486	0.122

把各个指标总体的平均值作为标准，对各类公司的指标进行打分。优秀值表示行业的最高水平，良好值表示行业的次高水平，中等值表示行业的中等偏上水平，平均值表示行业的总体平均水平，较低值表示行业的较低水平，较差值表示行业的最差水平。优秀值的类打分为100分，良好值的为80分，中等值的为60分，平均值的为40分，较低值的为20分，较差值的为0分（资产负债率为反向指标，打分则相反）。例如在流动比率指标中，最高的为类别6，所以类别6为优秀值，打分为100，其他依次为：类别2为良好，打分为80；类别4为中等，打分为60，类别1为平均，打分为40；类别5为较低值，打分为20；类别3为较差值，打分为0。其中，流动比率和资产负债率反映了偿债能力，应收账款周转率和总资产周转率反映营运能力，资产报酬率和净资产收益率反映盈利能力，净利润增长率和总资产增长率反映了成长能力。所以表2从偿债、营运、盈利和成长能力来进行综合打分，如偿债能力的评价结果就是通过对流动比率指标进行打分并乘以权重值和对资产负载率指标进行打分并乘以该指标的权重值相加得到的。其结果如表2所示。

表2 综合评价结果

类别	偿债能力	营运能力	盈利能力	成长能力
1	6.4	5.4	0	7.2
2	20	5.4	30	7.2
3	7.2	14.4	33.6	16.2
4	15.6	7.2	34.8	9
5	6.8	9	14.4	0
6	10	12.6	13.2	14.4

从表2中可看出，第1类是最差的，成长、偿债、营运、盈利能力都是最差的；第2类偿债能力最好，盈利能力较好，营运和成长能力较差；第3类是最好的，各个方面都很好；第4类盈利能力最好，偿债能力较好，营运和成长能力一般；第5类成长能力最差，偿债和盈利能力较差，营运能力一般；第6类营运和成长能力较好，其他一般。表2的统计结果不够直观，应用可视化技术可使用户直接发现各个类指标的好坏，进而能对公司的状况快速地了解。

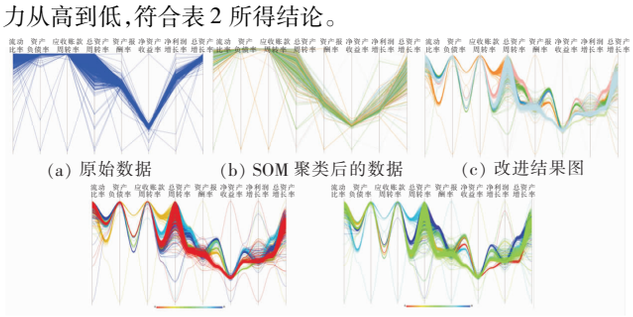
2 基于SOM的金融数据可视化

经过上述SOM聚类方法，本文对初始的金融数据进行了分类处理，并且获得了各个类别的系数。然而对于普通投资者来说，这样的分类结果仍然难以判断，不利于进一步的分析。对于金融分析师而言，对上述分类数据进行理解和分析亦是一项复杂而耗时的工作。因此，本文引入平行坐标和圆形平行坐标的可视化思想，设计了可视化系统，对分类后的数据进行可视化。可视化的结果大大提高了初始金融数据的分析效率，增强了可视性和可分析性。

2.1 平行坐标的可视化

平行坐标的基本思想是将 n 维数据属性空间通过 n 条等距离的平行轴映射到二维平面上，每一条轴线代表一个属性维，轴线上的取值范围从对应属性的最小值到最大值均匀分布。这样，每一个数据项都可以根据其属性值用一条折线段在 n 条平行轴上表示出来，相似的对象就具有相似的折线走向趋势。

SOM聚类后的金融数据在平行坐标中的显示如图2所示。图2(a)是未进行处理的原始数据，可以看到所有多维数据各个属性上的值，但是杂乱无章，很难理解这些数据，对于用户来说，根本不可能选出财务状况好的公司，作出投资决策。图2(b)是对数据用SOM聚类后的结果，分成了六类，可以看出一些趋势和规律，但仍是错综复杂，各个类的数据交叉绘制出来，有个大概的轮廓，但是还是很难辨别。本文进一步改进，用曲线代替直线，使得属性相近的线靠拢聚集，如图2(c)所示，使用户能更清晰地看到数据的内在联系并快速地找到自己感兴趣的数据，能够很清楚地看到六个类的范围。用户可根据自己的需求把各类按照偿债、运营、盈利和成长能力进行排序。以盈利能力为例，结果如图2(d)所示。在图2(d)中设置了颜色条，从蓝色到红色由高到低排序。盈利能力由净资产收益率和总资产报酬率所决定。从图2(d)中观察净资产收益率和总资产报酬率这两个属性轴间的曲线，可以看到深蓝色所表示的那一类公司盈利能力最好，依次降序排列为浅蓝、浅绿、深绿、桔红，最差的为红色（见电子版）。这个结论与表2所得结论一致。再以成长能力为例，成长能力是由净利润增长率和总资产增长率决定。如图2(e)所示，净利润增长率和总资产增长率这两个属性轴间的曲线，从深蓝到红依次排列，表明成长能力从高到低，符合表2所得结论。



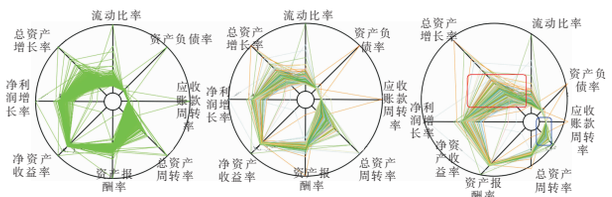
(a) 原始数据 (b) SOM聚类后的数据 (c) 改进结果图

(d) 盈利能力分类 (e) 成长能力分类

图2 平行坐标的显示

2.2　交互式圆形平行坐标的可视化

交互式圆形平行坐标是在圆中设置一个圆形的控制钮,属性轴以其为中心,向周围辐射延伸,并将圆均分,然后将平行坐标的坐标轴一一映射到这些线段上,把所有数据点也映射到这些线段上,这样可以形成一个封闭的圆形平行坐标图。圆形中可调节坐标位置的控制钮默认在圆心位置,用户可以根据自身的可视需求,自由移动控制钮,放大自己感兴趣的区域。比起图 2 所示的平行坐标,交互式圆形平行坐标使第一个属性轴和最后一个属性轴连接起来,能够更好地观察首尾坐标轴数据之间的关系。SOM 聚类后的金融数据在改进的圆形平行坐标中的显示如图 3 所示。



(a) 原始数据　(b) SOM 聚类后的数据　(c) 交互操作

图 3　交互式圆形平行坐标的显示

从图 3 可以看出,这种可视化结合了圆形和平行坐标两种方法。由于位于内部的低数据值和位于外部的高数据值具有明显的不对称性,通过这种可视化方式可以更容易地识别出确定的模式,可在不影响可视性和布局结构的情况下加入更多数据轴,适合多维数据。在交互式圆形平行坐标中,用户可以自由移动控制钮,放大自己感兴趣的区域。以图 3(c)为例,通过移动控制钮,红色框内的数据得到了放大,缩小了蓝色框内的数据。红色区域内的数据是净利润增长率和总资产增长率,所以,通过交互操作可以重点关注企业的成长能力。

3　结束语

本文提出了基于 SOM 聚类的金融数据可视化方法。首先根据金融数据的实际经济含义用 SOM 算法进行了分类,然后用平行坐标和交互的圆形平行坐标将多维数据在二维空间表示出来,从而使大量的多维金融数据能够同时被清晰地显示,便于发现数据的变化趋势及规律。

对于金融数据有很多的显示方法,但是传统的方法很难同时看到大量的信息。本文提出了用平行坐标和交互式的圆形平行坐标显示,从而使用户能同时看到大量公司的信息,减少了用户的工作量。另一方面,对于聚类,SOM 应用到金融数据的分类是合理和有效的。SOM 区别于其他技术方法的主要特点是自组织性和保序性,这样就保证了金融数据分类标准的客观性和特征提取的准确性。SOM 的模式识别功能又将模式相近的数据聚为一类。也就是说,同一类的金融数据具有相似的结构性特征,从而弥补了其他算法在发掘金融数据结构性特征上的不足。

在接下来的工作中,笔者计划进一步对平行坐标的可视化技术进行改进,使显示更加清晰,找出更多的数据隐藏的规律。在对金融数据的分析中,用户的经验是非常重要的,所以要进一步地完善各种交互,使用户的参与更多、分析更精准。在交互式圆形平行坐标中,类与类之间交错排列,杂乱无章,要进一步地在视觉上进行聚类,使各类能够清晰地区分开来。

参考文献:

[1] RICHARD M, PETER W M. Technical analysis of stock trends[M]. 9th ed. Boca Raton, FL: CRC Press, 2007.

[2] MURPHY J J. Technical analysis of the financial markets: a comprehensive guide to trading methods and applications[M]. New York: New York Institute of Finance, 1999.

[3] Dow Jones & Company Inc. Smart money map of the market[EB/OL]. http://www.smartmoney.com/marketmap/.

[4] HAVRE S, HETZLER B, NOWELL L. ThemeRiver: visualizing theme changes over time[C]//Proc of IEEE Symposium on Information Visualization. Washington DC: IEEE Computer Society, 2000: 115-123.

[5] ALSAKRAN J, ZHAO Ye, ZHAO Xin-lei. Tile-based parallel coordinates and its application in financial visualization[C]//Proc of the Visual and Data Analysis Conference at SPIE Electronic Imaging. 2010: 1-12.

[6] FAROOQ U, DILSHAD A, CHRIS N. Revealing trends of dynamic data in visualizations[C]//Proc of IEEE Symposium on Information Visualization. 2001: 8-11.

[7] SCHRECK T, TEKUŠOVÁ T, KOHLHAMMER J, *et al*. Trajectory-based visual analysis of large financial time series data[J]. ACM SIGKDD Explorations, 2007, 9(2): 30-37.

[8] TEKUSOVA T, KOHLHAMMER K. Applying animation to the visual analysis of financial time-dependent data[C]// Proc of the 11th International Conference on Information Visualization. Washington DC: IEEE Computer Society, 2007: 101-108.

[9] KEIM D A, NIETZSCHMANN T, SCHELWIES N, *et al*. A spectral visualization system for analyzing financial time series data[C]//Proc of Eurographics/IEEE TCVG Symposium on Visualization. 2006: 195-202.

[10] INSELBERG A, DIMSDALE B. Parallel coordinates: a tool for visualizing multi-dimensional geometry[C]//Proc of the 1st Conference on Visualization. LoS Alamitos, CA: IEEE Computer Society Press, 1990: 361-378.

[11] HOFFMAN P E. Table visualizations: a formal model and its applications[D]. Lowell: University of Massachusetts, 1999.

[12] WISHART D. K-means clustering with outlier detection[C]//Proc of the 25th Annual Conference of the German Classification Society. 2001: 14-16.

[13] DAS N. Hedge fund classification using K-means clustering method[C]//Proc of the 9th International Conference on Computing in Economics and Finance. 2003: 1-27.

[14] KRISHNAPURAM R, KELLER J M. A possibilistic approach to clustering[J]. IEEE Trans on Fuzzy Systems, 1993, 1(2): 98-110.

[15] 刘晓勇,林健良. 模糊聚类分析在金融机构财务分析中的应用[J]. 科学技术与工程, 2007, 7(1): 99-101.

[16] KAASTRA I, BOYD M. Designing a neural network for forecasting financial and economic time series[J]. Neurocomputing, 1996, 10(3): 215-236.

[17] ZHOU Yi-xin, ZHANG Jie. Stock data analysis based on BP neural network[C]//Proc of the 2nd International Conference on Communication Software and Networks. Washington DC: IEEE Computer Society, 2010: 396-399.

[18] KIM K J. Financial time series forecasting using support vector machines[J]. Neurocomputing, 2003, 55(1-2): 307-319.

(下转第 1306 页)

记为 r)、查准率(precision,简记为 p)。如果用列联表对单类赋值分类器的性能进行统计计算,则有两种方法可进行,即宏观平均和微观平均。宏观平均是先对每一个类统计 r、p 值,然后对所有的类求 r、p 的平均值;微观平均是先建立一个全局列联表,然后根据全局列联表计算。最终分类结果的好坏在一定程度上反映了特征选择的效果,因此可以通过评价分类结果来测试特征选择方法的效果。

根据文献[10],微平均精确率(micro-averaging precision)被广泛用于交叉验证比较。这里用它来比较不同的特征选择算法的效果。

3.3 实验步骤与结果

对数据集中的文本进行预处理,包括去停用词、过滤标点符号、进行词频和文档频率统计等。然后对数据集采用 10 折的交叉验证方法(10 fold cross validation),即将数据集随机划分成 10 个不相交的自己,进行 10 次训练和测试,依次将其中的一个子集作为测试集,其他子集作为训练集,取 10 次训练的平均值作为最终的分类结果。采用全局选取的特征选取方式,选取的特征数目设定为 1 000,特征加权方法为 TF × IDF,分类方法分别为 KNN 和 SVM。KNN 算法的 K-近邻值设为 35,对传统 CHI、IG 和新提出的 Var-CHI 方法进行评估。

图 1 显示的是 KNN 分类器分别采用 IG、CHI 和 Var-CHI 三种特征选择方法在 ICTCLAS 发布的中文数据集上的 micro_P 曲线。从图中可以看出,Var-CHI 方法在特征数目为 1 000 时,将 KNN 分类器的 micro_P 值从 86% 提高到 90%;Var-CHI 方法在选取 400 个特征时就可以使该分类器的 micro_P 值达到 89.61%;CHI 方法在选取 1 500 个特征时使该分类器的 micro_P 值达到 87.90%;IG 方法能够使该分类器的 micro_P 值达到最高值 87.69%。

图 2 显示的是 SVM 分类器分别采用 IG、CHI 和 Var-CHI 三种特征选择方法在 ICTCLAS 发布的中文数据集上的 micro_P 曲线。从图中可以看出,Var-CHI 方法在特征数目为 1 000 时将 SVM 分类器的 micro_P 值从 92% 提高到 96%;Var-CHI 方法在选取 400 个特征时就可以使该分类器的 micro_P 值达到 93.15%,并且在选取 1 000 个特征时使该分类器的 micro_P 值达到最高值后趋于稳定;CHI 方法在选取 1 300 个特征时使该分类器的 micro_P 值达到 93.15%;IG 方法能够使该分类器的 micro_P 值达到最高值 92.82%。

4 结束语

通过分析特征词与类别之间的相关性,在原有的卡方特征选择上增加三个调节参数:第一个参数通过计算特征词在某一文档中出现的频率解决 CHI 对低频词不可靠(夸大低频词的作用)的问题;第二个参数度量某一类特征词的文档频与所有类该特征词的文档频的平均值之间的偏离程度;第三个参数度量在特定类的某一文档中特征词的词频与这个类中所有文档词频的平均值之间的偏离程度。分别利用 KNN 和 SVM 分类算法进行对比实验,结果表明,基于方差的卡方特征选择方法使得查全率和查准率都得到了明显的提高。

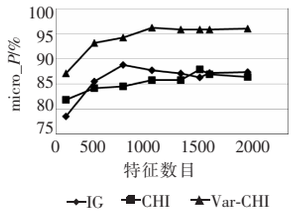


图 1 采用三种不同特征选择方法的 KNN 分类器性能比较

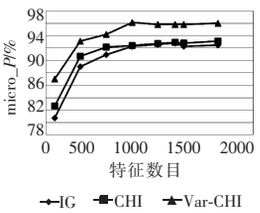


图 2 采用三种不同特征选择方法的 SVM 分类器性能比较

参考文献:

[1] 唐焕玲,孙建涛,陆玉昌. 文本分类中结合评估函数的 TEF-WA 权值调整技术[J]. 计算机研究与发展,2005,42(1):47-53.

[2] 苏金树,张博锋. 徐昕. 一种快速文本归类算法的设计与实现[J]. 软件学报,2006,17(9):1848-1859.

[3] 张莉,孙钢,郭军. 基于 K-均值聚类的无监督的特征选择方法[J]. 计算机应用研究, 2005,22(3):23-24,42.

[4] 周宇,覃征. 聚类分析中特征选择的研究[J]. 计算机应用研究, 2006,23(5):55-57,62.

[5] 杨恢先,杨心力,曾金芳,等. 在线特征选择的目标跟踪[J]. 计算机应用研究, 2010,27(3):1180-1182.

[6] YANG Yi-ming, LIU Xin. A re-examination of text categorization methods[C]//Proc of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:ACM,1999:42-49.

[7] 李粤,李星,刘辉,等. 一种改进的文本网页分类特征选择方法[J]. 计算机应用,2004,24(7):119-121.

[8] 张俊丽. 文本分类中的关键技术研究[D]. 武汉:华中师范大学, 2008.

[9] 余俊英. 文本分类中特征选择方法的研究[D]. 南昌:江西师范大学,2007.

[10] YANG Yi-ming. An evaluation of statistical approaches to text categorization[J]. Information Retrieval,1999,1(1-2):69-90.

[11] 李荣陆. 文本分类及其相关技术研究[D]. 上海:复旦大学,2005.

(上接第 1303 页)

[19] SHIN K S,LEE T S,KIM H J. An application of support vector machines in bankruptcy prediction model[J]. Expert Systems with Applications,2005,28(1):127-135.

[20] KOHONEN T. Self-organizing maps[M]. New York:Springer,1995.

[21] EKLUND T,BACK B,VANHARANTA H,*et al*. Assessing the feasibility of self-organizing maps for data mining financial information[C]// Proc of the 10th European Conference on Information Systems. 2002:528-540.

[22] SILVA B L, MARQUES N. Feature clustering with self-organizing maps and an application to financial time-series portfolio selection[C]//Proc of International Conference on Neural Computation. 2010:301-309.

[23] CARLOS S C. Self organizing neural networks for financial diagnosis[J]. Decision Support Systems,1996,17(3):227-238.

[24] 荆新,王化成,刘俊彦. 财务管理学[M]. 4 版. 北京:中国人民大学出版社,2006.

[25] MIZUNO H,KOSAKA M,YAJIMA H,*et al*. Application of neural network to technical analysis of stock market prediction[J]. Studies in Informatic and Control,1998,7(2):111-120.

[26] 孙宝珩,藏洪阁,张以宽. 现代商业企业财务分析与评价[M]. 北京:中国国际广播出版社,1996.