

聚类在股票研究中的应用

张迎春, 陈洁, 张晨希, 万忠, 张燕平

(安徽大学人工智能研究所, 安徽合肥 230039)

摘要:聚类是按照事物的某些属性, 把其聚集成类, 使各类间的相似性尽量小, 类内相似性尽量大。现在使用的一些聚类算法大多效率不高、聚类速度慢。文中在改进 LBG 算法的基础上提出了一种新的聚类算法, 克服了传统的 LBG 算法的缺点, 具有准确性高、测试时间短的优点。现将它应用于股票数据的预测分析中, 实验结果表明这种新的聚类算法, 相较于其它聚类算法能够取得更好的结果。

关键词:聚类; 股票; 预测

中图分类号: TP301.6

文献标识码: A

文章编号: 1005-3751(2006)04-0116-03

Application of Clustering in Study of Stock

ZHANG Ying-chun, CHEN Jie, ZHANG Chen-xi, WAN Zhong, ZHANG Yan-ping

(Institute of Artificial Intelligence, Anhui University, Hefei 230039, China)

Abstract: Clustering assembles things according to some of their attributes, minimize the comparability among clusters and maximize the comparability inside each cluster. Many algorithms now in used have some defects such as inefficient and slow. This text forwards a new arithmetic based on the improvement of LBG arithmetic. This new algorithm has gotten over the shortcomings of the traditional LBG with high veracity and needs short test time. Now it applied on stock data forecast analysis. The experiment shows this new clustering algorithm will yield better outcome than the old ones.

Key words: clustering; stock; forecast

通常说“物以类聚”, 就是说将性质相似的事物分成同一类, 所以在聚类分析^[1]中, 首先要给出有关论域中各元素之间的相似性的一个度量(如相似度函数), 然后给出一个聚类的原则, 根据其相似性和所给的聚类原则(最优原则)进行聚类, 寻求某种意义下的最优的解。在股票市场中对于众多上市公司的股票走势, 广大股民们很难做出正确的投资判断。如果具有关于上市公司的详细信息并且使用聚类算法进行分析, 那么将会做出正确的预测。文中主要介绍一种聚类方法在股票研究中的应用, 对现有历史的数据使用聚类算法进行聚类分析, 然后对结果中的异动点进行分析。

1 聚类的介绍

这里先介绍一下聚类。所谓的类就是指具有某些相

似元素的集合。所谓的聚类^[1]就是按照事物的某些属性, 把事物聚集成类, 使类间的相似性尽量小, 类内相似性尽量大。聚类(cluster)就是把一个 N 维的欧氏实时空间划分为 M 个区域, 这区域分别由其中心矢量表示。这个过程需要一个有大量的矢量构成的样本集, 通过统计试验后得出 M 个中心矢量, 这一过程叫做训练也就是我们所说的聚类过程。这 M 个中心矢量通常称为一个大小为 M 的码本, 每个中心矢量都成为一个码字。聚类分析的内容非常丰富, 常用的聚类方法^[2,3]有系统聚类法、传递背包法、动态聚类法、 C -均值聚类法等。聚类有一个重要的问题就是如何确定一个准则, 使得在这个准则下聚类过程达到最优, 也就是用这 M 个中心矢量可以最好地表示这个样本集。设有样本集 $X = \{x_i\}, 1 \leq i \leq N$, 把它聚成 M 类,

$$X = C_1 \cup C_2 \cup \dots \cup C_m (C_i \cap C_j = \emptyset, \text{当 } i \neq j)$$

这个分类记作为 C , 而其准则度量(距离)记作 $D(C)$ 。聚类的任务就是对 X 作一个最准分类 $C = C_1 \cup C_2 \cup \dots \cup C_m$, 使得 $D(C) = \min D(C)$ 。

一个理想的准则度量^[4]必须在主观上是有意义的而且易于运算, 这种度量有好多种, 常用的几种有:

$$D_1(C) = \text{tr}(w)$$

$$D_2(C) = \text{tr}(w^{-1}B)$$

$$D_3(C) = \text{tr}(T^{-1}B)$$

$$D_4(C) = \det W$$

收稿日期: 2005-07-13

基金项目: 国家自然科学基金重点基金资助项目(60135010); 安徽省自然科学基金资助项目(050420208); “九七三”计划(国家重点基础研究)资助项目(2004CB318108)

作者简介: 张迎春(1982-), 男, 安徽砀山人, 硕士研究生, 研究方向为人工神经网络、人工智能在金融工程中的应用; 张燕平, 教授, 博士, 研究方向为人工神经网络、机器学习、人工智能在金融工程中的应用。

其中 D_1 最常用,称为最小平方距离准则。

聚类的算法很多,比较常见的是 LBG 算法^[5],这种算法是非常经典的算法,聚类的结果还是比较好的,而且它的算法复杂度不高。但是它也有一些不足之处,例如它的结果过多依赖于初始向量是否分散从而使得出的结果不具有代表性。文中使用的是一种比较新颖的算法,它具有 LBG 算法的思想但是同时又引入了欧几里得距离的概念,克服了 LBG 算法的一些缺点,聚类结果比较准确。

下面介绍一下该算法的步骤:

1) 设定迭代次数 $L = 1000$, N 表示被测试的记录数, M 表示要聚的类数,很明显有 $M \leq N$ 。I. num 表示类 I 中的样本记录的数目。初始化 M 个中心点作为聚类参数,并且设置 $D_0 = 1$, $D = 1e + 10$, $m = 0$ 。

2) 当 $m < L$ 且 $\text{fabs}(D_0 - D)/D_0 > 1e - 5$ 时,则进行以下步骤。

3) 对于每个类设置相关参数:

$D_0 = D$, $D = 0$, $m++$

4) 对于所有被测试的样本记录,定义一个 $\text{Distance}[j] = 1e + 10$,然后定义一个 Dist 作为样本数据与聚类中心点的距离的平方和。分别比较 Dist 和 $\text{Distance}[j]$ 。如果 $\text{Dist} < \text{Distance}[j]$,则 $\text{Distance}[j] = \text{Dist}$,类 I 的记录数加一。 $D = D + \text{Distance}[j]$ 。

5) 如果类 I 的记录数为 0,则重新调整聚类中心点的数据大小。如果类 I 的记录数不为 0 则定义一个 M 表示类中记录最多的一个类,并且设记录的数目为 $M.\text{num}$ 。对于所有的样本记录比较它所在的类 I 是否与 M 相等,如果不相等则继续比较下一记录样本数据,如果相等则把 I. num 与 $(M.\text{num}/2)$ 做比较,如果 $I.\text{num} < (M.\text{num}/2)$,把当前的样本记录放到类 I 中,重新调整类 I 的相关参数, $I.\text{num}++$;如果 $I.\text{num} > (M.\text{num}/2)$,把当前的样本记录放到类 M 中,重新调整类 M 的相关参数, $M.\text{num}++$ 。若 $I.\text{num} > M.\text{num}$,则分别调整两者的相关参数。

6) 返回聚类的结果。

2 结果分析

笔者有众多上市公司近几年的财务报表,由于报表的数据比较庞大,直接进行分析是非常困难和复杂的。然而,进行聚类分析后数据就简捷多了,这就有利于做出快速而且准确的分析。首先,统计出所需要的有关公司的信息。它包括:每股收益、每股净资产、每股股利、净资产收益率、股利支付率、流动比率、负债比率、存货周转率、投资报酬率等一共 16 个属性。考虑到股票的价格波动主要受每股收益的影响,因此把每股收益作为主要属性。然后,按照不同的年份对报表进行分类,同一年份的作为同一类。其次,文中用某一年的数据进行聚类,分析聚类的结果。在聚类中把同一个数据分别分 3,4,5 类进行,综合比较,得出较好的结果。

例如,有关 1998 年的 911 家上市公司的聚类结果如

表 1 所示。

表 1 1998 年部分股票聚类结果

聚类数	异动点	所属行业	所在地址
3	01 = 1000782	纺织制造业	广东
	02 = 2600096	化学制造业	云南
4	01 = 1000782	纺织制造业	广东
	02 = 2600096	化学制造业	云南
5	01 = 1000782	纺织制造业	广东
	02 = 2600096	化学制造业	云南

这里的聚类结果比较好,很明显有两支股票始终都是在一起的。这里着重对这两支股票进行分析,它们的股票代码分别是 1000782 和 2600096。表 2 列出了这两个公司的近几年的一些数据。

表 2 1998 年异类股票的部分信息

股票代码	上市年份	净资产收益率	净值报酬率	流动比率	每股净资产	每股股利	每股收益
1000782	1995	0.100	0.374	1.044	4.530	0.182	0.453
1000782	1996	0.136	0.558	1.159	4.635	0.226	0.630
1000782	1997	0.105	0.526	1.688	7.983	0.000	0.839
1000782	1998	0.060	0.161	2.071	5.330	0.000	0.323
2600096	1995	1.397	1.397	2.267	0.266	0.000	0.371
2600096	1996	0.452	0.452	1.325	0.984	0.000	0.444
2600096	1997	0.193	0.372	6.763	2.345	0.415	0.451
2600096	1998	0.233	0.467	2.645	2.010	0.390	0.467

股票代码为 1000782 的公司是纺织企业,位于广东省。从 1995 年到 1998 年的历史数据发现该公司的净资产收益率和净值报酬率等在这两年有下滑趋势,这表明它的资产利用或成本控制发生了问题,估计其次年的每股收益将会下降。

股票代码为 2600096 的公司是从事化学原料及化学制品制造的,位于云南,是一家历史的老牌企业。它的净资产收益率、每股股利和每股净资产等有轻微的下跌幅度。但是它具有良好的流动比率和速动比率,又考虑到企业的性质,预计其下一年每股收益将会有轻微的下降。显然上述的是各种类型的公司在一起进行聚类的结果,除此之外,还可以对相同部门的公司进行聚类分析。例如,文中对 2000 年信息业的 64 家上市公司的数据进行聚类。由于这次聚类的总数比较少,分 4 类进行聚类时异动点数 2 比较理想,这两支股票的信息如表 3 所示。

表 3 2000 年信息业部分股票聚类结果

聚类数	异动点	所属行业	所在地址
4	01 = 1000014	计算机设备制造业	深圳
	02 = 2600083	电子元件制造业	成都

可以发现代码为 1000014 和 2600083 的股票为较少的一类,下面从历史数据进行分析。历史数据如表 4 所示。

表 4 2000 年信息业异类股票的部分信息

股票代码	上市年份	每股净资产	净资产收益率	存货周转率	每股收益
1000014	1998	1.330	-0.592	2.992	-0.787
1000014	1999	-0.131	6.359	3.782	-0.840
1000014	2000	0.002	-175.7	3.497	-0.373
2600083	1998	-1.217	-1.185	0.844	-1.442
2600083	1999	0.187	-4.538	0.522	-0.849
2600083	2000	0.120	-2.549	0.363	-0.307

股票代码为 1000014 的公司位于深圳市,从事计算机及相关设备的制造方面的业务;股票代码为 2600083 的公司位于成都,主营业务是电子元件的制造。它们从事的都是比较新型的产业,前两年都有着比较大的收益,可是近几年一直都在亏损。虽然每股收益一直是负的,但是它们也都在缓慢上升,尤其是股票代码为 1000014 的公司的存货周转率一直都比较高。因计算机正在中国广泛普及,所以它应具有较大的回升的潜力。股票代码为 2600083 的公司从 1998 年到 2000 年每股收益都有一个较大的回升,而且其各项数据稳中有升,估计每股收益还会再上升。

3 结束语

聚类方法对于处理大量的数据信息是很有效的。聚类分析综合众多上市公司的多项数据指标,能够较真实地

反映公司的盈利能力和水平,并且从中挖掘出实用的数据信息,根据综合的数据可以对上市公司的前景进行较准确的预测。通过对文中的测试结果以及算法的分析得出此方法在股票的预测中具有一定的应用前景。但是,仍有一些待改进的地方,目前正在研究中,力争使得该预测思想能够在现实中进一步推广使用。

参考文献:

- [1] 罗可,蔡碧野,吴一凡,等.数据挖掘中聚类研究[J].计算机工程与应用,2003(20):182-184.
- [2] 丁学钧,杨克俭,李虹,等.数据挖掘中聚类算法的比较研究[J].河北建筑工程学院学报,2004,22(3):125-127.
- [3] Berthold M R, Wiswedel B, Patterson D E. Neighborfram clustering Interactive exploration of cluster neighborhoods [A]. IEEE International Conference on data mining (ICDM'02)[C]. Maebashi City:[s.n.],2002.
- [4] Alexandros N, Yannis T, Yannis M. C²P: clustering based on closest pairs[A]. In: Apers P M G, Atzeni P, Ceri S, et al. Proceedings of the 27th International conference on Very Large Data Bases[C]. Roma: Morgan Kaufmann Publishers, 2001. 331-340.
- [5] 李滔,王俊普,吴秀清,等.基于改进的 lbg 算法的 SVM 学习策略[J].复旦学报(自然科学版),2004,43(5):789-792.

(上接第 106 页)

此方面的研究都围绕着在满足应用的 QoS 约束前提下如何最小化多播树费用问题上,极少有算法考虑到网络负载均衡分布问题。负载在网络中均匀分布的情况将直接关系到网络能否有效避免链路拥塞现象的发生,能否充分、合理地利用网络资源,从而提高网络运行性能。另外,笔者发现,大多数提出的路由算法采用无向图作为网络的数学模型^[5],即假设网络中两节点间方向相反的两条链路上的费用及延时等相等,这在实际网络中是不存在的。

(3)基于成组多播的路由算法问题。文献[6]针对 QoS 约束的成组多播路由问题提出一种构建多棵路由树算法,而其在求解过程中没有考虑路由树的费用问题和网络负载均衡分布问题。

(4)目前针对同时考虑网络节点的度约束和应用的 QoS 约束的多播路由算法的研究很少,仅文献[7]中提出一种基于延时和度约束的算法。因此,如何在算法中实现 QoS 约束,网络节点的度约束及负载均衡分布有效结合是有待解决的问题。

4 结束语

开发有 QoS 保证的基于 IPv6 的下一代因特网一直是近几年研究的热点问题。对下一代因特网的 QoS 路由机制的研究,能有效地解决网络拥塞与资源利用率问题,能

尽可能为用户提供可靠的网络服务并保证服务质量,也是将来网络发展要解决的关键问题。围绕该问题展开的各项研究工作,将有利于各种网络服务的发展和应用。

参考文献:

- [1] Kurose J F, Ross K W. Computer Networking: A Top - Down Approach Featuring the Internet[M]. 北京:人民邮电出版社,2004.
- [2] 薛希俊,孙雨耕,刘振肖.基于带宽和跳数的流量工程动态路由选择算法研究[J].电子学报,2002,30(2):274-278.
- [3] Wang Z, Crowcroft J. Quality of service routing for supporting multimedia applications[J]. IEEE Journal on Selected Areas in Communications, 1996, 14(7): 1288-1294.
- [4] 金明晔,李乐民,徐政五.一种基于 MPLS 业务量工程的选路机制[J].电子科技大学学报,2002,11(1):1-6.
- [5] Salama H F, Reeves D S, Viniotis Y. Evaluation of multicast routing algorithms for real - time communication on high - speed networks[J]. IEEE Journal on Selected Areas in Communications, 1997, 15(3):332-345.
- [6] Low C P, Song X Y. On finding feasible solutions for the delay constrained group multicast routing problem [J]. IEEE Transactions on Computers, 2002, 51(5): 581-588.
- [7] 刘莹,刘三阳,吴建平.多媒体通信的多播路由算法[J].电子与信息学报,2002, 24(7):948-953.