文章编号:1672 - 058X(2012)06 - 0050 - 07

数据挖掘在劳动力成本估计中的应用*

彭 茜[1],何淑明[2]
(1. 西南大学 经济管理学院,重庆 400700;2. 重庆工商大学 管理学院,重庆 400067)

摘 要:企业家在进行投资前考虑的一个重要方面是企业未来的盈利能力,而最终是否盈利则取决于收入和支出。员工的工资总额(劳动力成本的主要部分)作为企业的一项重要支出自然纳入企业家投资决策的考虑范围,对我国不同地区不同行业劳动力成本的合理估计对决策的制定有着重要意义。通过基于聚类分析的数据挖掘技术,结合实证分析对 19 个主要行业和 31 个省市自治区(中国大陆)的薪资水平进行了分析和对比,以更直观的方式揭示内在含义,为企业投资决策提供支撑。

关键词:数据挖掘技术;劳动力成本;聚类分析
中图分类号:C931.1 **文献标志码**:A

0 引 言

在经济全球化的大趋势下,企业的竞争最终会转化为人力资源之间的竞争,企业的发展对人力资源的依赖程度日益加深,怎样吸引、留住人才成为企业管理者的重要问题。通常较高的薪资水平成为引人、留人的基本筹码,但无疑增加了企业的成本,理性的企业家都会权衡收入和支出,并使其差值达到最大值,其结果决定了企业的盈利水平。所以,企业进行投资决策时,对不同地区、不同行业劳动力成本的合理估计显得尤为重要。

1 数据挖掘

1.1 数据挖掘的概念

数据挖掘的概念是 1995 年在美国计算机学会 ACM 会议上首次被提出的。它是从大量的、不完全的、有噪声的、模糊的原始数据中抽取隐含的,以前未知的,潜在有用的信息和知识的过程(W. J. Frawley, G. Piatetsky Shaprio)[1]。

数据挖掘是一门交叉性学科,它涉及人工智能、数据库技术、机器学习、模式识别、信息学、信息检索、统计学等多个领域[2-3]。在对数据库技术研究的历程中,相继出现了一些相似的术语,例如数据库中的知识发现(KDD)、数据融合(Data Fusion)等。

收稿日期:2012 - 01 - 18;**修回日期**:2012 - 02 - 18.
* **基金项目**:2011 年教育部人文社科青年基金项目"模型与实证:新生代农民工择业行为研究"(11XJC630005).
作者简介:彭茜(1987-),女,四川绵阳人,硕士研究生,从事人力资源管理研究.

1.2　数据挖掘在人力资源管理中应用的现状

数据挖掘在商业中的应用是很广泛的,最典型的例子是沃尔玛通过数据挖掘在大量数据中挖掘分析出小孩的尿布和啤酒之间有着惊人的联系。从大量的事实中得知该方法在挖掘已有数据中隐含的规律以及解决具体问题方面,是其他技术方法所不能比拟的。

因此,一些学者把其应用在人力资源管理上,分析与处理企业员工的相关信息。国外的研究有:Berry & Linoff (1997)[4]把数据挖掘技术应用在企业的市场销售预测和客户管理中;SAS 公司开发的数据挖掘工具 Enterprise Miner,用于企业在数据挖掘方面的应用和人力资源管理的决策支持应用;IBM 公司所开发的 Intelligent Miner 是一个客户服务系统,支持数据挖掘在人力资源管理中的应用。国内的数据挖掘技术主要集中在计算机领域,而在人力资源管理领域的研究主要集中在绩效评估和薪酬满意度的评估。而从劳动力成本角度进行研究的却比较缺乏,此处将基于此角度进行研究,探讨其对于企业家投资决策的作用。

2　研究过程

2.1　数据来源和方法

数据来源于2010 年《中国国家统计年鉴》,查到中国 31 个省市自治区(中国大陆)19 个主要行业的员工年平均工资(城镇单位),导入 SPSS18.0 中。利用公式 $Y_{ij}=\frac{x_{ij}-x_{\min}}{x_{\max}-x_{\min}}$ 将数据进行 0 - 1 标准化,再进行聚类分析,作树状图、散点图、柱状图等。

2.2　聚类分析原理和方法

聚类分析是数据挖掘的主要任务之一,它能按照事物的某些属性,把事物聚集成类,使类间的相似性尽可能小,类内相似性尽可能大。

聚类分析方法主要包括两种,即层次聚类分析方法(Hierarchical Cluster Analysis)和快速聚类分析方法(K-Means Cluster Analysis)。为了保证类别划分的准确性,可以用前一种方法进行分析,用后一种方法进行验证。

2.3　分析过程

劳动力成本包括职工工资总额、社会保险费、职工福利费、职工培训费、劳动保护费、职工住房费和其他成本费用。其中,职工工资总额是劳动力成本的主要部分,所以此处用员工年平均工资来评估劳动力成本水平。把 31 个省市自治区、19 个行业的城镇单位就业员工的年平均工资导入 SPSS18.0 进行聚类分析,可以得出相应的树状图和散点图。然后可以根据各图的特征获得投资者所需的信息,其作用主要体现在如下几点:

对全国劳动力成本的总体水平进行估计。决策者对全国劳动力成本总体水平进行宏观把握,是对劳动力成本合理估计的第一步。聚类分析则可以综合各个行业的劳动力成本,将劳动力成本相近的区域归为一类,有利于决策者了解各类别劳动力成本水平以及不同类别之间劳动力成本的差距。

按地区分类,对同行业各地区劳动力成本进行估计。决策者拟投资某一行业,但是需要考虑投资区域,此时采用的聚类方式是行业不变,地区聚类。树状图和散点图将年平均工资相近的地区划分为一类。投资

者可以首先以各个类别为整体进行分析,比较类别之间的工资差距,考虑企业的竞争力、资金实力、政治环境等因素,确定可以承受的劳动力成本后选择投资的类别,再就该类别的各个地区进行对比,结合企业家的个人偏好、资源、政策文化环境等因素,最后确定投资地区。

按行业进行分类,对同地区各行业劳动力成本进行估计。如果某个决策者想在某个地区进行投资,需要对投资的行业进行决策。软件的统计结果将很直观地得到19个行业大致的劳动力成本,而聚类分析将对劳动力成本接近的行业进行归类。以此为依据再结合企业实力并对比该地区综合劳动力成本水平做出投资决策。

2.4 实证过程

此处涉及的城镇主要的19个行业,主要为:农、林、牧、渔业,采矿业,制造业,电力、燃气及水的生产和供应业,建筑业,交通运输、仓储和邮政业,信息传输、计算机服务和软件业,批发和零售业,住宿和餐饮业,金融业,房地产业,租赁和商务服务业,科学研究、技术服务和地质勘查业,水利、环境和公共设施管理业,居民服务和其他服务业,教育,卫生、社会保障和社会福利业,文化、体育和娱乐业,公共管理和社会组织。

2.4.1 全国各地各行业综合劳动力成本分析

从图1,图2可以清楚得出,综合19个行业的薪资水平,上海和北京的劳动力成本最高,划为第一类,人均劳动力成本为5.6~5.8万/年;其次是天津、西藏、广东、江苏和浙江也比较高,划为第二类,人均劳动力成本为3.6~4.6万/年;剩下的各个地区的劳动力成本相对较低的,划归为第三类,人均劳动力成本为2.5~3万/年。每一类别人均劳动力成本的差距在1.3~1.6万/年。结合实际可以看出,聚类分析的结果和实际的劳动力成本和地区经济发展状况是紧密相接的,但是也有特殊的例子,比如:西藏地区的总体经济情况不及第三类地区,但是它划归为第二类地区。其中的原因是,此处研究对象是城镇19个行业员工的薪资水平,而不是整个地区的所有人的收入情况,而就西藏城镇水平来讲由于享受国家的政策倾斜和地理位置等原因使其劳动力成本较高,将其划为第二类地区是无可厚非的。

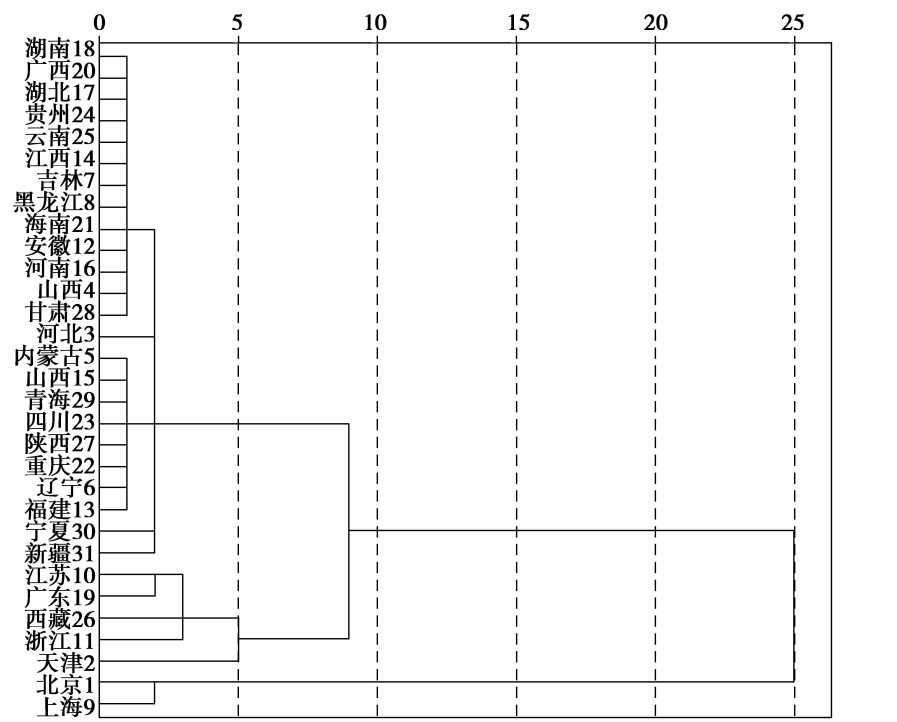

图1 各行业综合树状图

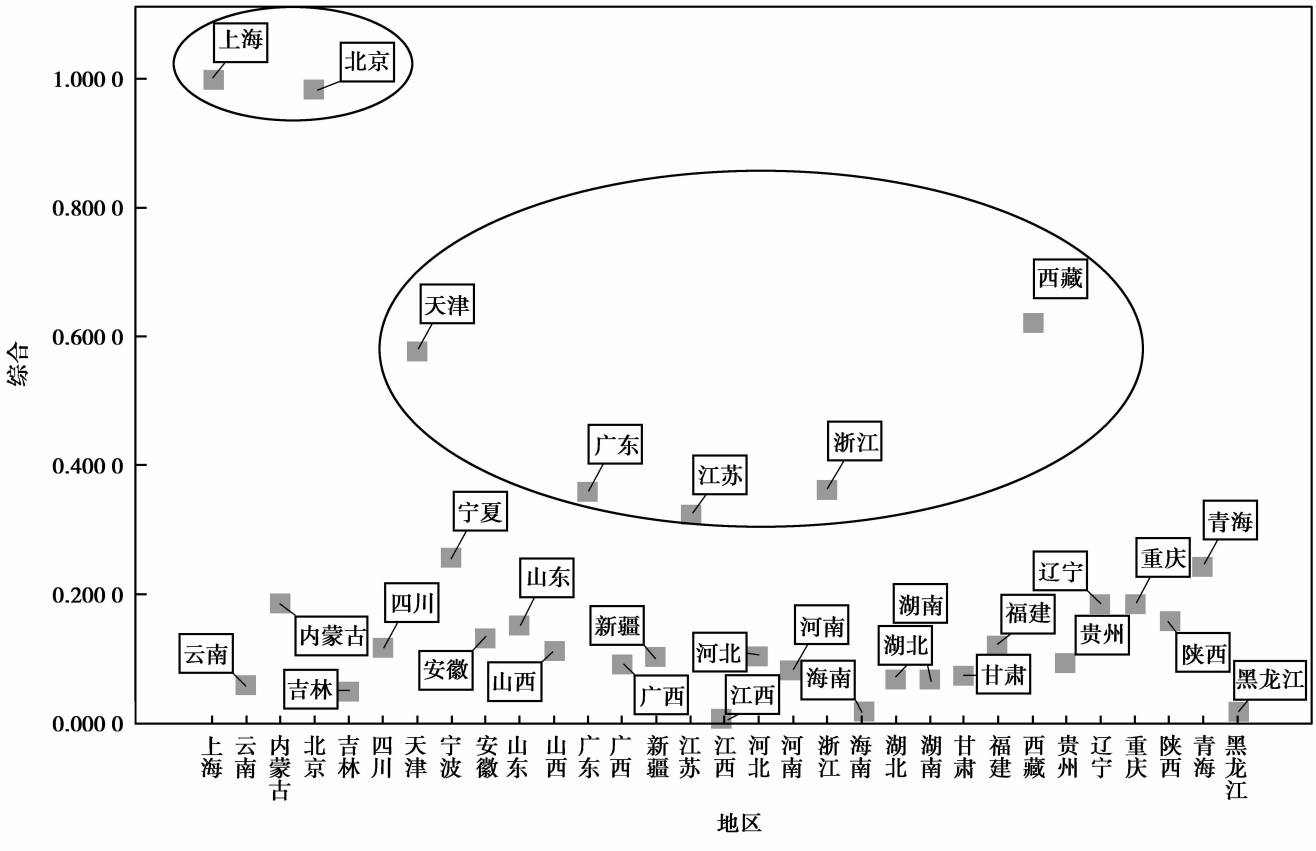

图 2　各行业综合水平散点图

上面例子是采用层次聚类进行分析,可以用快速聚类验证其分类的合理性。

首先,把划分的类别设定为 3 类,得出的结果如下(表 1),可以看出,该方法划分的每一类的地区和层次聚类的结果是一致的。

表 1　快速聚类表

case number	地区	Cluster	Distance	case number	地区	Cluster	Distance
1	北京	1	0.412	17	湖北	3	0.315
2	天津	2	0.972	18	湖南	3	0.402
3	河北	3	0.589	19	广东	2	0.617
4	山西	3	0.600	20	广西	3	0.355
5	内蒙古	3	0.483	21	海南	3	0.507
6	辽宁	3	0.499	22	重庆	3	0.505
7	吉林	3	0.483	23	四川	3	0.377
8	黑龙江	3	0.391	24	贵州	3	0.380
9	上海	1	0.412	25	云南	3	0.510
10	江苏	2	0.525	26	西藏	2	0.781
11	浙江	2	0.658	27	陕西	3	0.395
12	安徽	3	0.527	28	甘肃	3	0.500
13	福建	3	0.639	29	青海	3	0.700
14	江西	3	0.511	30	宁夏	3	0.678
15	山东	3	0.536	31	新疆	3	0.560
16	河南	3	0.329				

随后,可以根据方差的结果对划分为3类的正确性进行检验,分析简略后的结果如下表(表2)。从表中可以看出,在分为3类的前提下,各个行业中除“居民服务业和其他服务业”外,其他的行业对应的相伴概率都小于0.01的显著水平,说明根据该种划分,各个类别的地区之间存在显著的差别。而对于居民服务业和其他服务业,3个类别之间的 F 统计量的相伴概率为0.03,小于显著水平0.05,认为居民服务业和其他服务业在不同类别地区中存在比较显著的差异。所以,从方差的检验结果可以验证层次聚类分析的可靠性和正确性。

表2 各行业聚类方差表(ANOVA)

行业	F	Sig	行业	F	Sig
农、林、牧、渔业	16.924	.000	房地产业	56.808	.000
采矿业	6.371	.005	租赁和商务服务业	50.558	.000
制造业	60.699	.000	科学研究、技术服务和地质勘查业	57.334	.000
电力、燃气及水的生产和供应业	64.018	.000	水利、环境和公共设施管理业	54.212	.000
建筑业	38.269	.000	居民服务和其他服务业	3.984	.030
交通运输、仓储和邮政业	23.141	.000	教育	49.698	.000
信息传输、计算机服务和软件业	61.035	.000	卫生、社会保障和社会福利业	91.335	.000
批发和零售业	117.311	.000	文化、体育和娱乐业	111.836	.000
住宿和餐饮业	69.786	.000	公共管理和社会组织	105.314	.000
金融业	152.845	.000			

2.4.2 针对某一行业对不同地区的劳动力成本进行分析(以房地产行业为例)

假设某房地产公司欲在某个地区设立分公司,现在考虑投资的地区,在此忽略政策等其他因素的影响,

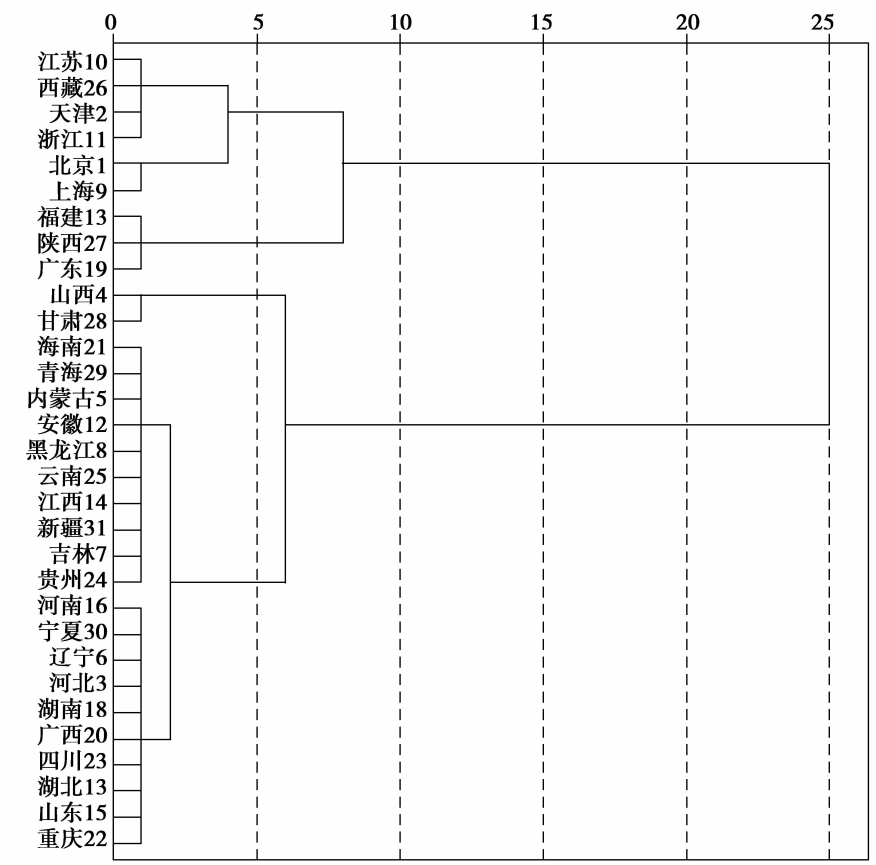

图3 房地产行业树状图

主要把劳动力成本纳入考虑的范围。就此,从图 3 的分析结果可以得出,全国从事房地产行业的人均劳动力成本大致可以分为 4 类,第一类的平均劳动力成本最高,包括上海、北京、天津、江苏、浙江和西藏,人均劳动力成本 3.8 ~4.5 万/年;第二类的劳动力成本次之,主要包括广东、福建、陕西,人均劳动力成本 3.1 ~3.4 万/年;第 4 类是劳动力成本比较低的地区,包括山西和甘肃,人均劳动力成本 1.6 ~1.8 万/年;第三类的劳动力成本处于中等水平 2.0 ~2.8 万/年,即为余下的地区。

事实证明,房地产行业各个地区的劳动力成本差距小于行业综合水平,尤其是前三类地区劳动力成本差距偏小。而房地产行业的销售价格则和当地经济发展有着密切的关系,决策者就偏向对经济发展程度高的地区进行投资,这就解释了经济发达地区房地产行业泡沫多的原因。当今中国房地产价格走势处于不明朗时期,劳动力成本呈上升趋势,并且此行业存在较大的进入和退出门槛,所以投资决策分析显得尤为重要。决策前需要充分考虑各种因素,不仅要考虑企业的劳动力成本,还需要考虑国家政策、经济发展程度、需求量等因素。

2.4.3　针对某一地区对不同行业劳动力成本进行分析(以重庆为例)

同理,对重庆 19 个行业进行层次聚类分析,可以得出重庆地区劳动力成本最高的几个行业是信息传输、计算机服务和软件业,金融业,科学研究、技术服务和地质勘查业和电力、燃气及水的生产和供应业;其次是教育,公共管理和社会组织,卫生、社会保障和社会福利业;其他的行业则相对较低。劳动力成本高的行业主要分布在国家垄断和科技含量较高的行业,对投资者的资质要求较高同时也有较高的利益回报,投资者在进行决策前一定要把握好国家的宏观政策,正确估计自己的实力,合理评估投资风险。而从政府角度来讲,为了促进上述高劳动力成本行业更好的发展,应鼓励企业家投资并逐步放开某些行业的投资要求,给予投资者适当的补贴和税收优惠政策。

3　结　论

通过以上实证研究,将数据挖掘技术中的聚类分析运用于劳动力成本估计,能更迅速、准确地比较出不同地区不同行业的劳动力成本差距,这为企业投资决策提供重要的事实依据。

但是,此处的研究还存在一些局限的地方。首先,样本只是 31 个省市自治区,对劳动力成本的估计只能停留在市级及以上行政地区的总体估计,而市级以下的行政区域无法精确估计,这对决策的准确程度会有影响。第二,文中进行了一些假设,如忽略政策因素、文化冲突、商业禁忌、家庭背景等因素,但这些因素在实际决策过程中起到一定的影响作用。第三,把员工工资总额作为企业劳动力成本,而忽略了如社会保险费、职工福利费等成本,有待进一步研究。

参考文献:

[1] HAN J I,MICHELINE K. 数据挖掘——概念与技术[M]. 北京:高等教育出版社,2001

[2] 施蕾,孟凡荣. 数据挖掘系统结构的研究[J]. 微计算机信息,2007,6(3):167 - 168

[3] 薛静. 基于时间序列算法与多层次分布式智能决策支持系统[J]. 计算机工程与设计,2007,8(2):3646 - 3664

[4] BERRY M J,LINOFF G. Data mining techniques:For marketing,sales,and customer support[M]. John Wiley & Sons,1997

[5] 杨辉. 数据挖掘分类优化方法研究[D]. 上海:上海交通大学,1999

[6] 王明宇. 基于数据挖掘的人员配置模型研究[D]. 北京:北京林业大学,2010

Application of Data Mining to Labor Cost Estimation

PENG Qian[1], HE Shu-ming[2]

(1. School of Economics and Management, Southwest University, Chongqing 400715, China;
2. School of Management, Chongqing Technology and Business University, Chongqing 400067, China)

Abstract: An entrepreneur emphatically considers the future profitability of the enterprise before conducting investment, however, whether the enterprise has the profit depends on earnings and expenditure, the total sum of employees wages (main part of labor cost), as an important expenditure of the enterprise, is naturally considered by the entrepreneurs' investment decision-making, as a result, the rational estimation on labor cost in different regions and in different industries of China plays an important role in investment decision – making. Based on data mining of clustering analysis, through empirical study, this paper makes analysis and comparison of wages level in 19 industries and 31 provinces and municipalities of mainland China and provides support for enterprise investors by more intuitively revealing the inherent meaning of the estimation.

Key words: data mining technique; labor cost; clustering analysis

责任编辑:代小红
校　　对:李翠薇

(上接第49页)

Multiplication Algorithm on Binary Field Based on NAF_w

JIANG Hong-bo[1], WU Yan[1], FENG Xin-yu[1], DU Yan-qiu[1], YANG Qing-jiang[1], SHI Ke-ying[2], LIU Yan-wei[2]

(1. College of Electric and Information Engineering, Heilongjiang Institute of Science and Technology, Harbin 150027, China; 2. Heilongjiang University, Harbin 150001, China)

Abstract: The speed of multiplication on elliptic curves is a key to improving performance of Elliptic Curve Cryptography(ECC). This paper analyzes the non-adjacent form (NAF) algorithm of the width w and the polynomial multiplication algorithm and proposes a multiplication algorithm on binary field based on NAF_w. This algorithm reduces the XOR operation in the frequency and the number of precomputation, decreasing the computation time and saving storage space. The modeling and simulation results show that its average efficiency is approximate 14.7% faster than the comb polynomial multiplication and it only needs $2^{w-1}-1$ precomputation, Based on the number of precomputation to storage and time consumption, $w=4$ is better choice to the width of the window.

Key words: elliptic curve; NAF; binary field

责任编辑:代小红