Available online at www.sciencedirect.com
SCIENCE DIRECT®
Data & Knowledge Engineering 53 (2005) 311–337
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

Linear correlation discovery in databases: a data mining approach

Roger H.L. Chiang [a,*], Chua Eng Huang Cecil [b], Ee-Peng Lim [c]

[a] *Information Systems Department, College of Business, University of Cincinnati, P.O. Box 210211, Cincinnati, OH 45221-0211, USA*
[b] *Nanyang Business School, Nanyang Technological University, Singapore 639798, Singapore*
[c] *Center for Advanced Information Systems, School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore*

Received 8 September 2004; accepted 8 September 2004
Available online 19 October 2004

Abstract

Very little research in knowledge discovery has studied how to incorporate statistical methods to automate linear correlation discovery (LCD). We present an automatic LCD methodology that adopts statistical measurement functions to discover correlations from databases' attributes. Our methodology automatically pairs attribute groups having potential linear correlations, measures the linear correlation of each pair of attribute groups, and confirms the discovered correlation. The methodology is evaluated in two sets of experiments. The results demonstrate the methodology's ability to facilitate linear correlation discovery for databases with a large amount of data.

Keywords: Knowledge discovery in database; Linear correlation; Association measurement; Data mining

* Corresponding author. Tel.: +1 513 556 7086; fax: +1 513 556 4891.
E-mail address: roger.chiang@uc.edu (R.H.L. Chiang).

0169-023X/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2004.09.002

1. Introduction

As competition among businesses continue to increase, it is crucial for organizations to discover knowledge that could give them advantages over their competitors. In the past, such knowledge often was obtained by collecting data and testing it against some predefined hypothesis (i.e., a hypothetico-deductive approach to obtaining knowledge). Lately, greater emphasis has been given to discovering (inducing) knowledge from existing databases. The knowledge discovery approach employs various data mining algorithms such as association rule mining algorithms [1] to obtain knowledge from databases. However, very little work has investigated the possibility of automating traditional data analysis using statistical methods for knowledge discovery [2–5]. Because the amount of data generated and accumulated continues to exceed the number of available experienced analysts [6], it is imperative to develop methods to automate and expedite data analysis for knowledge discovery from existing databases [7,8]. This research establishes a novel discovery methodology to induce business knowledge (also called business intelligence) in the form of linear correlations for better decision making.

As an illustration of the usefulness of such automated knowledge discovery, consider an organization with globally distributed factories that wants to determine how factory effectiveness can be improved. Factory effectiveness includes various aspects, such as, cost per unit produced, output per day and factory downtime. Furthermore, many possible factors could influence factory effectiveness, including wages, reliability of supply, and age of the factory. In traditional data analysis, data analysts must first propose a set of hypotheses for testing. Statistics software, such as SAS/STAT and SPSS Base, can provide only the mechanisms to test these possible relationships [9]. Therefore, data analysts are responsible for ascertaining the appropriate analysis that will identify relationships through hypothesis testing, and must manually select the appropriate factor, outcome, and measurement function for each analysis. Often, fatigue, overhead, manpower cost, and the limits of human cognitive capability detract from a thorough understanding and complete analysis of the sheer volume of data available from existing business databases.

In this research, we demonstrate that these manual tasks of traditional data analysis (i.e., proposing hypotheses and selecting factors, outcomes, and measurement functions) can be automated for knowledge discovery in databases. Specifically, we automate linear correlation discovery (LCD), the goal of which is to determine whether two attributes or sets of attributes (i.e., attribute groups) have a relationship. A thorough discussion of LCD is presented in Section 2.1.

Some previous work has addressed related problems. Hou [3] developed a system that determined whether a regression or a classifier were appropriate for analyzing a system. The SNOUT project [2] derived some properties of attributes that could be leveraged for analysis. Aladwani [4] created an expert system to select an appropriate multiple-comparison test. Some authors have developed clustering or classification algorithms based on correlation (e.g., derivatives of Principle Component Analysis [5]), while others have developed pre-processors to select the 'best' algorithm for a given task [7,8]. However, to our knowledge, no one has attempted specifically to automate linear correlation discovery.