應用資料探勘技術於遺漏值問題之研究

The Study of Data Mining Technologies for the Problem of Missing Value

李仁鐘
華梵大學資訊管理系教授
johnlee@hfu.edu.tw

陳璽安
華梵大學資訊管理系研究生
m9725003@cat.hfu.edu.tw

摘要

資料探勘(Data Mining)已經是許多企業普遍使用的技術。在從紙本資料轉換成電子資料的過程中有可能因為人為疏失或是資料老舊等外在因素造成資料遺漏，往往遇到此種情形的作法都是將該筆資料刪除，或是以平均數、0、眾數等方法填補其遺漏值(Missing Value)，但此種作法在資料筆數少的情況下，一定會影響其資料的正確率，終將無法提供可靠的資訊給使用者。

本研究以公開資料集做為測試，隨機選取了裡面數筆資料的值當作遺漏值，並用平均數、0、及支援向量迴歸(Support Vector Regression, SVR)分析之數值回填，最後用迴歸樹(Regression Tree)來分析比較，結果顯示使用支援向量迴歸所預測出的數值遺漏值正確率最高。

關鍵詞：資料探勘、遺漏值、支援向量迴歸、迴歸樹。

Abstract

Data Mining is now widespread used for many enterprises. There could be missing data from paperwork to electronic system because of human error or out-of-date information. Usually these data might be deleted or using average value, 0 and mode value to fill the missing values, but this can only applicable for fewer data. It will certainly affect the accuracy of data and ultimately unable to provide reliable information to the user.

In this study, we use open datasets in our test. We use some data with missing values at random from the open datasets, then use average value, 0 and Support Vector Regression (SVR) to analyze numerical backfill. Finally we use regression tree to analyze the comparisons. The result shows that anticipation value by using SVR has the highest accuracy for missing value.

Keywords: Data Mining, Missing Value, Support Vector Regression, Regression Tree.

1. 研究動機與目的

針對過去有許多學者提出對遺漏值問題解決之方法，但其眾多方法之中，遺漏值回復率並不是盡善盡美，本研究希望藉由數個公開資料集，以資料探勘方式找出遺漏值回復率最高的方法，以便在未來遇到有遺漏值之資料，可以給使用者提高其資料的參考價值。

本研究以公開資料集作為測試，隨機選取數筆資料當遺漏值，再以 0、平均數、支援向量迴歸(Support Vector Regression, SVR)預測之數值回填，最後用迴歸樹方法計算出原始值及其他三種方法的平均誤差(Average Error)，比較其正確率，期能找到一套遺漏值回復率最高的方法。

2. 文獻探討

本研究將針對遺漏值及所使用資料探勘之技術，包括：支援向量迴歸、迴歸樹此兩種方法來進行探討。

2.1 遺漏值(Missing Value)

在 1950 年，就已經有學者提出有關遺漏值的文獻，然而，在面對大量資料，或者是在將紙本資料轉換為電子檔案的過程之中，一定會有人為疏失等等外在因素導致資料缺失，其遺漏的資料大略分為以下數種型態(楊棋全，2004)：

(一)空值(Empty Value)

此資料型態可能發生於問卷調查，在公開類型的問卷中，受訪者往往不是每個問題都填入答案，基於個人因素某些問題則不予以作答，此時回收回來的資料就會有空數值的產生。

(二)不存在值(Nonexistent Value)

會出現此型態的情況，是因為根本沒有該資料可填入。例如要預測某家公司股票未來的走勢，也許需要從五年前到現在的資料來分析預測，但是這家公司是兩年前才成為上市公司，所以前三年的資料就成為了不存在值。

(三)不完整資料(Incomplete Data)

舉例來說，要分析土壤是否適合種植植物，需要使用儀器分析土壤裡的成分，但是也許因為人為對自然環境的破壞，導致於機器無法偵測到某塊土壤裡的成分，或是因為人為疏失，機器操作不當等因素，造成了不完整資料。

(四)未蒐集到之資料(Uncollected Data)

此資料是指沒有去搜集到的來源，例如在訪談中沒有錄音，沒有錄影存證，沒有去留下任何紀錄可以給未來作為參考之用的，就會成為未蒐集到之資料(林俊男，2005)。

以上是大部分會在資料中遇到的遺漏值型態，但是在使用者參考資料遇到這些遺漏值時，大多數的處理方式有以下幾種：

(一)忽略該筆資料

忽略其資料遺漏值的部份，照常拿來做分析或參考，此種作法在參考價值上會有一定的偏差

(二)刪除該筆資料

直接把包含遺漏值的該筆資料刪除，此種作法對於資料量大的時候或許偏差值不高，但是面對資料量小，筆數少的資料的時候，在抽樣上很可能會產生誤差，進而降低其資料的參考價值。

(三)使用眾數取代其遺漏值

將該屬性出現最多次的數填補其遺漏值，但此種方法準確度相對來說會較低。

(四)使用平均數取代其遺漏值

為現在遇到遺漏值時最常使用之方法。求出該屬性的平均值後，再回填至遺漏值，此種方法相對也是種有可能會產生偏差的方法(林俊男，2005)。

2.2 支援向量機(Support Vector Machine, SVM)

支援向量機是由 Vapnik 在 1995 年和 AT&T 實驗室團隊所提出的統計學習理論（Statistical Learning Theory）所發展出的學習演算法(Vapnik, 1995)。其原理是以核心運算為基礎，使用核心函數來處理資料不但容易以線性的方法處理非線性的分類問題，還可以避免

實際坐標做映射運算時遇到的大量複雜運算，這樣一來便可透過統計學理論在特徵空間中尋找把訓練資料分類正確的最佳超平面(黃祺偉，2009)。

支援向量機的運用方法是以既有的案例作為訓練樣本，再利用這些分析出的資料(Training Data)選出支援向量(Support Vector)來表現整體的資料，並將少部分極端值事先剔除，然後將所挑的向量包裝成模型(Model)，再以此模型去判別所要分類的案例(陳偉明，2004)。

SVM 發展至今，其功能早期侷限在處理分類的問題，直到 1997 年，Vapnik 提出一個新想法稱做 Vapnik's ε-insensitive Loss Function，將 SVM 應用在迴歸問題上，才發展出支援向量迴歸(Support Vector Regression, SVR)(劉翔瑜，2006)。然而，支援向量迴歸與支援向量機最大的不同就是其輸出目標的部份，支援向量機主要是用來處理分類問題，而支援向量迴歸是用來處理迴歸問題，其兩者的輸出，前者是整數，後者是實數。其概念是一種基於非線性核心的迴歸模型，主要是要在高維度的特徵空間中找出擁有最小風險的迴歸超平面能夠準確預測資料的分佈，並希望能具有良好的函數逼近功能以及泛化能力(黃祺偉，2009)。

2.3 決策樹(Decision Tree)

在機器學習(Machine Learning)中，決策樹已經算是普遍被使用的工具，其顧名思義就是一個樹狀的結構，由樹根(Root)跟樹葉(Leaf)所組成，而中間從樹根到樹葉有許多節點(Node)或分叉點(Branching Point)，每一個葉子就代表一個最終的決策。在現有的決策樹分類器中，以 ID3 (Quinlan, 1986)、C4.5 (Quinlan, 1993)、CART (Breiman et al., 1984)最廣為採用。上面三種方法處理的問題主要為分類(Classification)問題，而所適用的資料型態是離散狀態或是文字符號。Quinlan 在 1996 年，改良了傳統 C4.5，使其在處理連續數值的資料型態時有了更進一步的方法，並發展出了迴歸樹(Regression Tree)。本研究採用的方法為 Quilan 所建立的迴歸樹，可處理的資料型態為連續值。該方法在概念上與決策樹相同，都是依據資料屬性進行分類，兩者差異處在於迴歸樹的終端節點是一條迴歸方程式，而不是如決策樹的終端節點是該筆資料所屬類別(陳樹衡，2007)。

迴歸樹將估計模型以決策樹的方式呈現，依據資料屬性，將資料分割成若干區塊，其過程會建立出許多規則(Rule)，再各自於區塊中選取重要的自變數，建立個別的迴歸模型最後計算出平均誤差(Average Error)，經由計算出來的誤差來判斷其正確率，其平均誤差的公式如下(陳樹衡，2007)：

$$\text{Average Error} = \frac{1}{n}\sum_{i=1}^{n}\left|d_i - y_i\right| \qquad (1)$$

其中 d_i 為實際值，y_i 為預測值，n 為資料總筆數。

3. 實驗方法與結果

3.1 研究資料

本研究是從 UCI Machine Learning Repository(http://archive.ics.uci.edu/ml/) 網站所提供的公開資料集作為研究資料，從網站中取兩個其型態為連續數值的資料當作測試，A 資料中有 506 筆，14 個屬性，B 資料中有 103 筆，10 個屬性，在這兩個資料當中，隨機抽取十筆資料，每筆資料中隨機抽取五個屬性當作其遺漏值，希望透過資料探勘技術，能夠得到可觀的遺漏值回復率。

3.2 實驗

本研究在遺漏值的部分採用 0、平均值與支援向量迴歸所分析之數值回填，最後以迴歸樹方法計算出原始值及其他三種方法的平均

誤差(林智仁，LibSVM)。

在 A 資料使用 0 當做其遺漏值後，用迴歸樹分析兩者的平均誤差並做比較，原始值平均誤差為 1.79，使用 0 當做遺漏值的平均誤差為 1.98，詳細的分析數值如圖 1 及圖 2 所示：

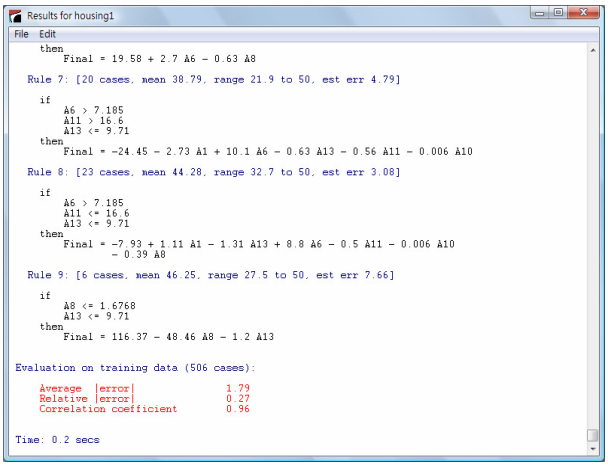

圖 1　A 資料使用迴歸樹分析資料原始值時所得的平均誤差

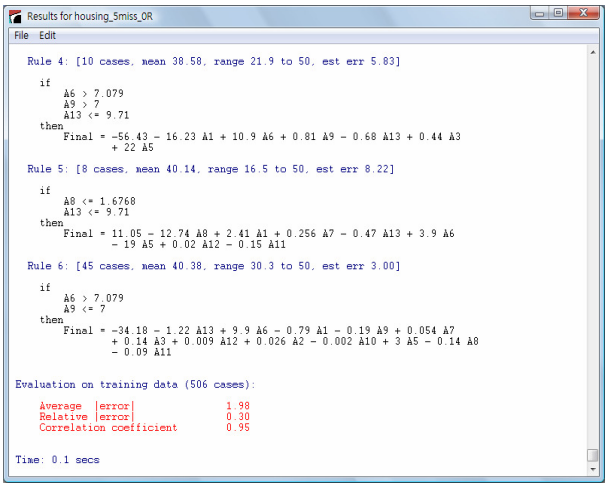

圖 2　A 資料使用迴歸樹分析以 0 當做其遺漏值時所得的平均誤差

在使用平均值當作其資料遺漏值時，所得的平均誤差為 1.84，詳細如下圖 3 所示：

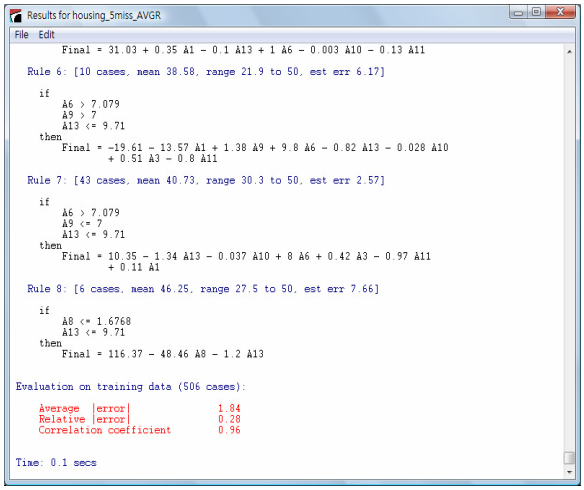

圖 3　A 資料使用迴歸樹分析以平均值當做其資料遺漏值時所得的平均誤差

在使用支援向量迴歸分析之數值當作其資料遺漏值時，所得的平均誤差為 1.79，詳細如下圖 4 所示：

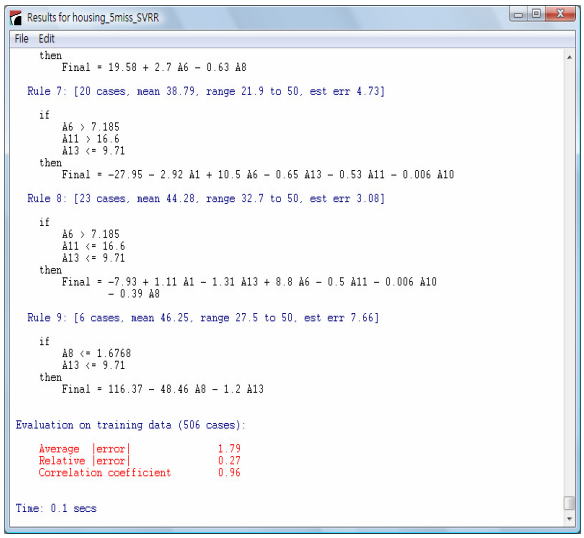

圖 4　A 資料使用迴歸樹分析以支援向量迴歸分析之數值當做其資料遺漏值時所得的平均誤差

下表 1 為 A 資料原始值與支援向量迴歸所分析數值之比較表：

表 1　A 資料原始值與支援向量迴歸分析數值比較表

		遺漏值1	遺漏值2	遺漏值3	遺漏值4	遺漏值5
第一筆	原始	78.9	242	17.8	396.9	9.14
資料	SVR	78.8	242.1	17.9	396.8	9.23
第二筆	原始	61.1	242	17.8	392.83	4.03
資料	SVR	61.1	242.1	17.9	392.73	4.12
第三筆	原始	45.8	222	18.7	394.63	2.94
資料	SVR	45.9	222.1	18.6	394.53	3.03
第四筆	原始	96.1	5.9505	311	396.9	19.15
資料	SVR	95.9	5.8503	311.1	396.8	19.05
第五筆	原始	100	6.0821	311	386.63	29.93
資料	SVR	99.8	5.9822	311.1	386.53	29.82
第六筆	原始	7.87	94.3	311	392.52	20.45
資料	SVR	7.96	94.1	311.1	392.42	20.34
第七筆	原始	29.3	307	21	386.85	6.58
資料	SVR	29.4	307.1	20.9	386.75	6.67
第八筆	原始	98.1	307	21	376.57	21.02
資料	SVR	98	307.1	20.8	376.47	20.92
第九筆	原始	100	307	21	394.54	19.88
資料	SVR	99.89	307.1	20.9	394.64	19.77
第十筆	原始	100	307	21	376.73	13.04
資料	SVR	99.8	307.1	20.8	376.63	13.14

以下為 B 資料，同樣也是以 0，平均值，支援向量迴歸分析之數值回填其遺漏值，再以迴歸樹計算出平均誤差，再與原始值比較，分析結果出來，原始值之平均誤差為 1.928，以 0 回填之平均誤差為 3.04，以平均值回填之平均誤差為 2.021，以支援向量迴歸回填之平均誤差為 1.929，詳細的分析數值如下圖 5678 所示：

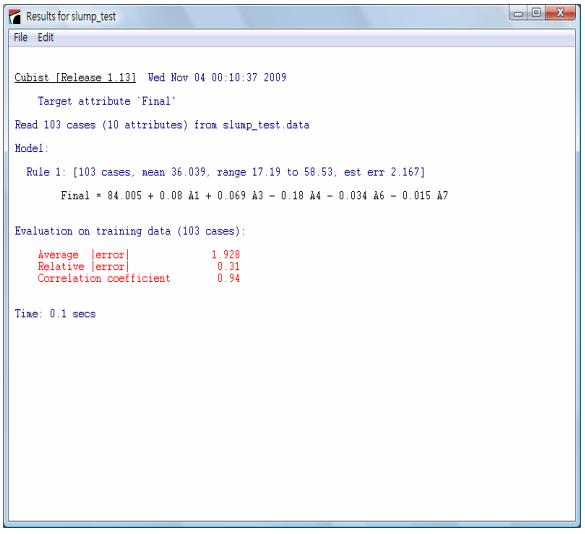

圖 5　B 資料使用迴歸樹分析資料原始值時所得的平均誤差

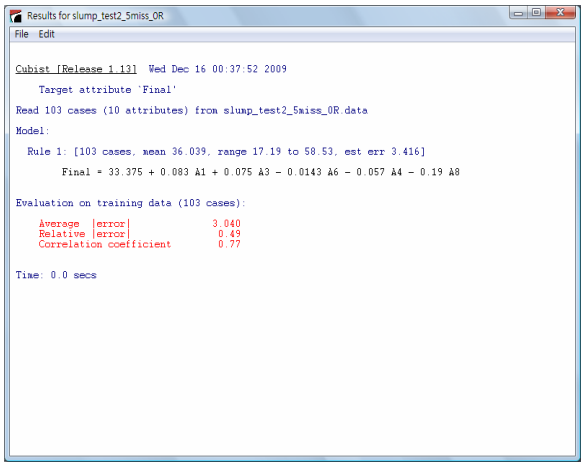

圖 6　B 資料使用迴歸樹分析以 0 當做其遺漏值時所得的平均誤差

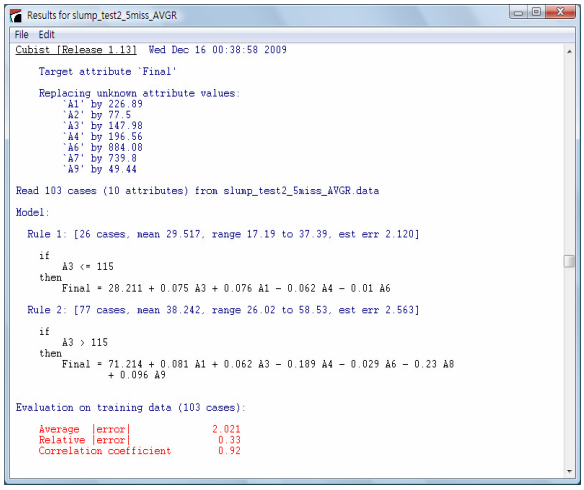

圖 7　B 資料使用迴歸樹分析以平均值當做其資料遺漏值時所得的平均誤差

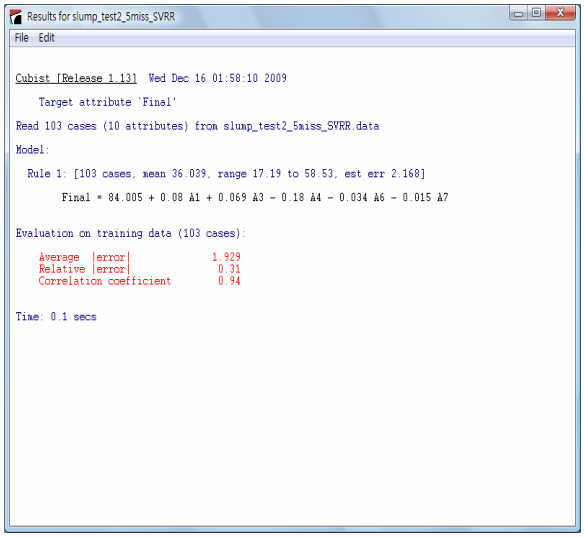

圖 8　B 資料使用迴歸樹分析以支援向量迴歸分析數值當做資料遺漏值時所得的平均誤差

下表 2 為 B 資料原始值與支援向量迴歸所分析數值之比較表：

表 2　B 資料原始值與支援向量迴歸分析數值比較表

		遺漏值1	遺漏值2	遺漏值3	遺漏值4	遺漏值5
第一筆資料	原始	273	82	210	904	680
	SVR	272.9	81.8	209.9	903.9	680.1
第二筆資料	原始	147	89	202	860	829
	SVR	147.1	88.9	201.9	860	828.9
第三筆資料	原始	153	239	200	1002	684
	SVR	153.1	238.9	199.9	1001.9	684.1
第四筆資料	原始	295	136	206	766	68.5
	SVR	294.9	136.1	205.9	765.9	68.4
第五筆資料	原始	296	219	932	685	48.5
	SVR	295.9	218.9	931.9	685.1	48.6
第六筆資料	原始	100	196	959	705	49
	SVR	99.9	196.1	958.9	705.1	49.1
第七筆資料	原始	295	136	208	871	650
	SVR	294.9	136.1	207.9	871.1	650.1
第八筆資料	原始	366	187	191	824	757
	SVR	365.9	186.9	191.1	824.1	756.9
第九筆資料	原始	274	89	202	759	827
	SVR	273.9	88.9	201.9	759.1	826.9
第十筆資料	原始	252	76	194	835	821
	SVR	251.9	76.1	194.1	835.1	820.9

下表 3 將 AB 資料的原始值及三種方法回填遺漏值之平均誤差做一個總整理。

表 3　AB 資料原始值及三種方法回填遺漏值之平均誤差比較表

	原始值	0	平均值	SVR
A資料平均誤差	1.79	1.98	1.84	1.79
B資料平均誤差	1.928	3.04	2.021	1.929

以上結果顯示，使用支援向量迴歸所分析出來之數值回填，其平均誤差與原始值最為貼近，其次是以平均值回填，最後是以 0 回填，表示所有方法之中，以支援向量迴歸所分析出來之數值其遺漏值回復率為最高。

3. 結論與未來研究方向

在本研究中使用了兩個公開資料做測試，並且使用三種方法填補資料遺漏值，分別用 0、平均值，支援向量迴歸分析之數值，再與原始值互相比較平均誤差，借此來判斷其回復率是否準確，A 資料原始值之平均誤差為 1.79，以 0 代入遺漏值之平均誤差為 1.98，以平均值代入其平均誤差為 1.84，以支援向量迴歸分析之數值代入其平均誤差為 1.79。B 資料原始值之平均誤差為 1.928，以 0 代入遺漏值之平均誤差為 3.04，以平均值代入其平均誤差為 2.021，以支援向量迴歸分析之數值代入其平均誤差為 1.929，實驗結果中發現以支援向量迴歸分析之數值回復率最高，其次是平均值，最後是 0，表示以目前的實驗來說，使用支援向量迴歸來做遺漏值之預測，可使其資料的參考價值提升至最高。

在未來，本研究將繼續以其他資料探勘技術做測試，例如倒傳遞類神經網路等技術分析遺漏值，期能找到比支援向量迴歸回復率更高，使用時間更短之方法，同時再多使用幾組資料做為測試，目的希望可以找到一種最適合針對解決遺漏值問題的方法。

參考文獻

[1]林俊男，”應用類神經網路法於遺漏值問題之研究”，***南華大學資訊管理學系碩士論文***，2005。

[2] 林智仁，LibSVM，URL: http://www.csie.ntu.edu.tw/~cjlin

[3]陳偉明，”以支援向量機改善船舶交通管理資訊系統之可疑動態目標偵測”，***華梵大學資訊管理系碩士論文***，2004。

[4]陳樹衡、郭子文、棗厥庸，”以決策樹之迴歸樹建構住宅價格模型—台灣地區之實證分析”，***住宅學報第十六卷第一期***，pp. 4-7，2001。

[5]黃祺偉，”多核心支援迴歸向量機應用於股價預測”，***南華大學資訊管理學系碩士論文***，2009。

[6]楊棋全，”指數與韋伯分佈遺失值之處理”，***國立中央大學統計研究所碩士論文***，2004。

[7]劉翔瑜，”倒傳遞類神經網路、支援向量迴歸於日經 225 現貨指數之預測及交易策略之研究”，***天主教輔仁大學金融研究所碩士論文***，2006。
[8]Breiman, L., J. H. Friedman, R. A. Olsen & C. J. Stone., “***Classification and Regression Trees***,” CA: Wadsworth, 1984.
[9]Quinlan, J. R., “Introduction of Decision Tree,” ***Machine Learning***. Vol. 1, pp. 81-106, 1986.
[10]Quinlan, J. R., “***C4.5: Programs for Machine Learning,***” CA: Morgan Kaufmann, 1993.
[11] Quinlan, J. R., “Improved use of continuous attributes in C4.5,” ***Journal of Artificial Intelligence Research***, Vol. 4, pp. 77–90, 1996.
[12]Vapnik, B. N., ***The Natural of Statistical Learning Theory***, 2nd Edition, Springer-Verlag, N.Y., USA, 1995.