**以決策樹為本的趨勢性離散化之研究**

|  |  |  |
| --- | --- | --- |
| 李御璽 | 顏秀珍 | 王品傑 |
| 銘傳大學資訊工程學系 | 銘傳大學資訊工程學系 | 銘傳大學資訊工程學系 |
| leeys@mail.mcu.edu.tw | sjyen@mail.mcu.edu.tw | pp8101313@hotmail.com |
|  |  |  |

**摘要**

 真實資料中包含許多數值型的資料；例如年齡、薪資、貸款期間等。數值型資料因範圍變化大且複雜度高，若直接使用會大大降低模型的穩定性。為了提升資料的穩定性並降低模型的複雜度，我們使用離散化(Discretization) 的技術，將數值型屬性的資料轉換成類別型屬性。然而，前人所提出的方法並無法將資料平均的分佈到每個區間，且只能找出線性的趨勢(遞增或遞減)。因此，本論文將會提出一個改進的方法，使得資料離散化後，能夠將資料平均地分佈到每個區間且找出非線性的趨勢，並同時保有原始資料的分佈趨勢。實驗的結果顯示，我們所提出的方法能有效地增加分類模型的F-measure，提升模型的效能。

**關鍵字 :** 決策樹、趨勢性離散化

**Abstract**

 In reality, data contain many numeric attributes like age, income and loan duration. Using numeric attributes to build a classification model would cause the model unstable because of the variety and complexity of numeric attributes. Therefore, we use discretization technique to reduce the variety and complexity of numeric attributes, which increases the importance of attributes. This study involves discretizing numerical attributes into categorical attributes in order to get a more accurate representation of the distribution with respect to the distribution of a target attribute. The trend discretization proposed previously has increased the magnitude of explanation and its accuracy. However, the amount of data contained in each interval is not distributed equally according to the overall intervals, and the algorithm can only detect linear trends. Therefore, we propose an enhanced trend discretization algorithm that not only equally distributes data into each interval, but it is able to detect curvilinear of the trend. The newly proposed algorithm can increase F-measure and model accuracy.

**Keywords :** Decision Tree, Trend Discretization

**壹、前言**

資料探勘(Data Mining)為從現有的大量資料中，擷取不明顯、之前未知且可能有用的資訊[10]。實作資料探勘的過程可分為三大部分，分別為資料前處理(Data Preprocessing)、資料探勘、和分析報告(Reporting)。

資料前處理包含了資料清理(Data Cleaning)、資料整合(Data Integration)、資料篩選(Data Selection)與資料轉換(Data Transform)，資料探勘則為演算法的應用。資料探勘的技術可分為四大類，分別為分類(Classification)、預測(Prediction)[1]、集群分析(Clustering)與關聯規則(Association Rules)，其中分類為最被廣泛運用的技術[9]，諸如影像分類(Image Classification)[7]、偵測惡意郵件(Virus detection of new malicious emails)[19]、銀行業中的風險評估與進件違約風險管理[18]。在整個資料探勘流程中，又以資料前處理為最重要的步驟，亦所謂使用高品質的資料才能獲得高品質的模型與結果[8]。

本論文將會使用德國信用卡資料集(German Credit Dataset) 為範例資料集，真實資料中包含了許多數值型屬性的資料，例如年齡、薪資、貸款期間等，而這種數值型資料因為變化大且複雜度高，若直接使用會大大降低整個模型的預測能力。舉例來說，，我們將資料集的70% 作為訓練資料來建立模型，30%作為測試資料來驗證模型的準確度，圖 1為數值屬性 Age 未經離散化之分佈狀況為年齡在未經離散化與違約機率的分佈，我們很明顯地發現訓練資料與測試資料均呈現高低起伏的分佈。



圖 1 數值屬性 Age 未經離散化之分佈狀況

 訓練資料與測試資料分佈差異很大容易造成模型過度學習(Over-fitting)，如果再次抽樣建模，會發現與第一次的模型長的非常不一樣，造成模型的不穩定。

 為了提升資料的重要度並降低複雜度，我們會使用離散化(Discretization)[17,2,3]的技術，將數值型屬性的資料轉換成為類別型屬性，如圖 2是將年齡離散化成5個區間，資料成現較平穩的狀態，訓練資料與測試資料呈現較一致的分佈，如此可以避免模型過度學習。

接下來第貳章將會介紹相關工作，包含分類與趨勢性離散化，第参章介紹我們的方法與第肆章實驗結果與比較。



圖 2 數值屬性Age經離散化後之分佈狀況

**貳、文獻探討**

**2.1 分類 (Classification)**

 決策樹最早是由Quinlan 於 1979 年所提出的ID3(Iterative Dichotomister) 演算法，其後又有C4.5 與 C5.0，這兩個演算法都是ID3的改進版本。

ID3演算法首先會先去計算在每個在S資料集裡的屬性的信息熵(Entropy)，信息熵是用來計算資料集裡的不確定性，若屬性的信息熵越低，代表該屬性的不確定性越低，我們就會選擇該屬性作為第一個用來切分資料集的屬性，我們可以將信息熵公式化如下:



S - 目前正在被計算信息熵的資料集

X - 在S中所有類別的集合

p(x) - x類別在S中所佔的比例

當信息熵越低，代表該資料集的不確定性越低，我們即可用該變數作為分割條件，將資料集一分為二。

**2.2 離散化 (Discretization)**

 早期離散化都是使用較簡單的技巧，如等寬裝箱法(Equal-width)和等深裝箱法(Equal-frequency)來作資料的離散化，但是隨者模型準確度與效率的需求提高，離散化的演算法也不斷地進步發展。現今許多離散化演算法已經被提出且證明可以降低資料的維度同時增加預測的準確性。我們可以五個面向來區分目前已提出的離散化演算法[15][24]，分別為:監督式(Supervised)與非監督式(Unsupervised)、靜態(Static)與動態(Dynamic)、全域(Global)與局域(Local)、由上而下(Top-down、Splitting)與由下而上(Bottom-up、Merging) 以及直接(Direct)與漸進式(Incremental)。

 離散化演算法可以是監督式或非監督式，主要區別是在離散化的過程中，是否有參考到目標屬性的資訊，監督式離散化會參考目標屬性的資訊如目標屬性在某區間的比例，相反之，非監督式演算法則不會參考目標屬性的資訊，如等寬裝箱法[5]或等深裝箱法[4]。

 我們亦可以以動態或靜態來區分，動態的方法會在建立分類器(Classifier)的過程中對數值型屬性進行離散化，相對地，靜態的方法會在建立分類其之前就先將數值型屬性離散化。

 另一面向為全域或局域，局域的方法只會使用部分資料去進行離散化，這種通常又是動態的；相反之，全域的方法會使用全部的資料去進行離散化，而這種離散化通常又是靜態的。

 離散化還可以以由上而下或由下而上來區分，由上而下亦稱為分割法，一開始所有資料都會落在同一個區間，然後逐步地分割區間後，再將所有資料重新分配給其所屬的區間；由下而上的方式又稱為合併法，一開始先將資料進行較粗略的分類，之後再藉由條件進行精細地合併區間，如ChiMerge[11]、Modified Chi2[22]、Extended Chi2[21]。

 最後一面向為直接與漸進式，直接的方法直接將資料切成使用者自訂的N個區間，如等寬與等深即為直接的離散化方法，至於漸進式的方法，一開始會產生一個較粗略的離散化結果，之後再逐步改進，並且需要一個終止條件或標準，來停止離散化。

 由於本篇所提出的方法為監督式離算化的一種，所以接下來我們將會以監督式與非監督式的面向來討論離散化，更深入地了解監督式與非監督式的區別。

**2.3 趨勢性離散化(Trend Discretization)**

 趨勢性離散化為2012年所提出的離散化方法[13]，該研究發現若屬性與目標屬性之間具有趨勢關係，亦為遞增或遞減，則將該屬性套用在羅吉斯迴歸中有較高的機率也會呈現具有趨勢性的迴歸方程式，如此一來，離散化的變數可以較容易地被解釋與應用。

 趨勢性離散化的目的較著重於應用面，也就是解釋能力，我們知道許多離散化的方法將數值型屬性離散化後，結果是無法解釋的，圖 3為使用等深離散化將貸款期間離散成5個區間後，違約機率的分佈狀況，這個離散化結果也許不錯，但是它無法被解釋為什麼要將貸款期間離散化成這幾個區間。



圖 3 等深離散化將貸款期間離散成5個區間

 而圖 4為貸款期間Duration經趨勢性離散化後的結果，可以很明顯的看出一個趨勢性，『當貸款期間越長，違約機率就越大』，這個離散化的結果不僅可以與實務經驗整合，還可以協助業務人員與建模人員去解釋離散化之結果，並了解產業知識 (Domain Knowledge)。



圖 4 貸款期間經趨勢性離散化後的結果

 接下來將會介紹整個趨勢性離散化的流程，並舉一個例子，更詳細的了解每一個步驟。趨勢離散化會先使用等寬裝箱法初步地對數值型屬性離散化，它預設每個區間的資料量不得低於 N/25，N為資料總筆數。然後計算每個區間內目標屬性出現的機率，計算方式如下:



 計算完每個區間的目標屬性出現機率後，利用這些機率值，我們求得一條簡單線性迴歸 (Simple Linear Regression)，y=∝x+β，在利用斜率∝來判斷該屬性在整體資料是呈現「遞增分佈」或是「遞減分佈」，遞增分佈亦指∝>0，反之遞減分佈∝<0。接下來將會在所產生的區間尋找切點，該切點必須使切割後的兩個區間有最大目標屬性出現機率的差，且其機率差必須大於0.01並保有趨勢性，找到最大機率差作為切點後，繼續在切開後的兩個資料集裡繼續找出個別最大切點，直到機率差小於0.01或未保有趨勢性，即完成趨勢性離散化，表 1為每一迴圈所找出的最佳切點。

表 1 趨勢性離散化之切點



圖 5 為貸款期間經趨勢性離散化後，與目標屬性-違約機率之分佈狀況，可以得到當貸款期間越長，客戶違約機率就越大。



圖 5 貸款期間經趨勢性離散化後的機率分佈

由我們可以很明顯地看出貸款期間與目標變數具有遞增的趨勢性，但是我們發現，每個區間所包含了資料量並不是趨向一個平均值如圖 6所示，第1個(BIN1\_9~15)和第2個區間(BIN2\_16~26)分別包含了34%的資料，但第3個區間卻只包含了6%的資料，這可能會造成我們所找出的趨勢性並不符合真實情況，亦或是偽趨勢。

 再者，我們發現趨勢性離散化只能找出單調的趨勢，亦指遞增或遞減，無法找出非單調趨勢，如果只是一味的將資料集離散成遞增或遞減，有些關鍵的年紀或數據的趨勢就無法找出。



圖 6 每個區間所包含的資料量

在第参章，我們將會提出一個改進的演算法，去改進舊演算法的兩個問題，第一，資料離散化後每個區間的資料量分佈不平均，第二，只有偵測單調性趨勢。

**参、研究方法**

 前些章節，我們了解了為什麼需要離散化並了解趨勢性離散化的目的與優點，趨勢性離散化是預期離散化後的結果可以被解釋，並同時提高模型的穩定度。但我們發現，趨勢性離散化有兩項不足的地方，第一，離散化後每個區間所包含的資料量並不平均，第二，趨勢性離散化只能找出單調性的趨勢，如遞增或遞減，無法找出非單調性的趨勢，如峰型或谷型。

本篇研究將會提出一個改進的演算法，更平均的將資料分佈到每一區間，除了可以找到遞增或遞減的趨勢，還可以偵測峰型與谷型的趨勢。

 我們先來舉個例子，介紹我們的演算法預期會得到甚麼結果，圖 7為年齡與違約風險原始資料之分佈，我們用之前的方法(趨勢性離散化)將其離散後，如圖 8所示，圖中的百分比數字代表該區間所包含的資料比例，我們算出其標準差為10.84，而圖 9為我們的方法將年齡離散化的結果，其標準差為9.45，表示資料分佈的更平均，且離散化後資料分佈的情況更接近原始資料的分佈。



圖 7 年齡與違約風險原始資料分佈



圖 8 之前的方法將年齡離散化之結果



圖 9 我們的方法將年齡離散化之結果

 首先我們先利用公式 1 皮爾森樣本相關係數找出樣本中屬性與目標變數間的相關係數，利用統計學的信心區間(Confidence Interval) 去推估出真實的相關係數 ρ，若ρ的lower bound與upper bound皆為小於0.4，則說明屬性與目標變數間不具有線性關係，反之，若ρ的lower bound 或upper bound其中一個大於0.4，則表示屬性與目標變數間可能具有線性關係。



公式 1 皮爾森樣本相關係數

之後，初步利用C4.5將數值離散化成數個區間，若屬性與目標變數之間具有線性關係，我們即會使用前演算法的方式，將區間進行趨勢性離散化(遞增或遞減)，反之，若不具有線性關係，則我們先將每個區間的數值由大到小排序，在兩兩數值間找到平均數做為切點，圖 10為一切點示意圖，圖中之橫線為其中一個切點，這個切點可以將0和1分為一群、2為一群、3和4為一群如圖 11所示，確保在離散化後的結果，能夠保有原來資料的分佈。



圖 10 切點示意圖



圖 11 以(圖10)中橫線為切點，離散化後結果

 在附上整個方法的完整流程圖如圖 12所示，首先我們先推估出真實的相關係數ρ，然後利用ρ的lower bound 與 upper bound 判斷屬性與目標變數間是否具有線性相關，接下來先進行C4.5初步將資料離散程數個區簡，接著若具有線性關係，則依照樣本的趨勢，進行遞增或遞減的合併，反之若未具有線性相關，則使用我們所提出的方法將資料進行合併。



圖 12 我們的方法完整流程圖

**肆、實驗結果**

 利用我們的方法將年齡屬性離散化後，我們得出每個區間裡所包含的資料量之標準差(9.45)比之前的方法所算出的標準差(10.8)還要小，說明資料分佈的更平均，且同時保有原始資料中屬性與目標變數間的趨勢性。



圖 13 我們的方法離散化後資料分佈量

 我們比較了數個資料集的數值型屬性，離散化後，比較標準差，如表2所示。平均下來，我們的方法(DTB)有效地將資料更平均的分佈至每個區間。

表 2 資料標準差比較表



 我們也比較了離散化後的變數，對於模型的影響程度，我們為每個變數建了一個多層感知機(Multilayer Perceptron)。首先我們在原始資料中，利用Information Gain找出前幾個重要變數，搭配離散化後的變數去建模，再利用10-fold validation去驗證模型，算出F-measure。

表 3 為利用幾個變數加上每種方法離散化後的變數，個別建模所算出的F measure，表中 Original 為未離散化的原始變數、CACC[24] 為近年做的較好的離散化方法[13]、TREND為之前演算法離散的方法[13]、DTB為本篇論文所提出的方法。左邊欄位是使用的資料集與被離散化的變數，方括號裡為參與建模的變數。

表 3 F measure 比較表

 表中可以看出我們的方法有效地提升F measure，增加模型預測準確度。

**伍、結論**

 適當地離散化數值型屬性，可以減少屬性的複雜度與雜訊，提升整體屬性的重要度，進而增加分類模型的準確度。

 本篇研究提出了一個改進的演算法，可以改進前演算法只能找出單調趨勢與資料分佈不均的問題，若變數與目標變數間具有線性關係，我們即以趨勢型之方式進行資料合併，找出趨勢關係；反之若變數與目標變數沒有線性關係，我們提出了一個合併的方法，可以保有原始資料的分佈趨勢。

 實驗結果顯示，我們的方式可以將資料更平均的分佈至每個區間，並可以增加模型的準確度。

**參考文獻**

1. G. Cardoso, F. Gomide, “Newspaper demand prediction and replacement model based on fuzzy clustering and rules”, Information Sciences, pp. 4799 - 4809, 2007.
2. J. Catlett, “On Changing Continuous Attributes into Ordered Discrete Attributes”, Proceedings European Working Session on Learning, pp. 164 - 178, 1991.
3. J.Y. Ching, A.K.C. Wong, K.C.C. Chan, “Class-dependent discretization for inductive learning from continuous and mixed mode data”, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 641 - 651, 1995.
4. D. Chiu, A. Wong, B. Cheung, “Information discovery through hierarchical maximum entropy discretization and synthesis ”, Knowledge Discovery in Databases, pp. 125 - 140, 1991.
5. P. Clark, T. Niblett, “The CN2 algorithm”, Machine Learning, pp. 261 - 283, 1989.
6. S. Cohen, L. Rokach, O. Maimon, “Decision-tree instance-space decomposition with grouped gain-ratio” in Information Sciences, pp. 3592 - 3612, 2007.
7. D. Go’mez, J. Montero, J. Ya’n’ez, “A coloring fuzzy graph approach for image classification”, Information Sciences, pp. 3645 - 3657, 2006.
8. P. Guo, S.S. Chen, Y. He, “Study on data preprocessing for daylight climate data” Proceedings of the Third International Conference on Information Computing and Applications, pp. 492 - 499, 2012.
9. G.K. Gupta, “Introduction to Data Mining with Case Studies”, 2006
10. J. Han, M. Kamber, “Data Mining: Concepts and Techniques, Morgan Kaufman”, 2001.
11. R. Kerber, “ChiMerge: Discretization of Numeric Attributes”, Proceeding of Ninth International Conference on Artificial Intelligence, pp. 123 - 128, 1992.
12. L. Kurgan, K.J. Cios, “CAIM discretization algorithm”, IEEE Transactions on Knowledge and Data Engineering, pp. 145 - 153, 2004.
13. Y.S. Lee, S.J. Yen and C. H. Lu, “A Discretization Algorithm Based on the Trend of Class-attribute Distribution”, Proceedings of International Symposiums on Major Academic Disciplines, 2012.
14. H. Liu, F. Hussain, C.L. Tan, M. Dash, “Discretization: an enabling technique”, Journal of Data Mining and Knowledge Discovery, pp. 393 - 423, 2002.
15. H. Liu, R. Setiono, “Feature selection via discretization”, IEEE Transactions on Knowledge and Data Engineering, pp. 642 - 645, 1997.
16. J.L. Lustgarten, V. Gopalakrishnan, H. Grover, S. Visweswaran, “Improving Classification Performance with Discretization on Biomedical Datasets”, Proceedings of AMIA Annual Symposium, pp. 45 - 449, 2008.
17. Z. Markzuki, F. Ahmad, “Data Mining Discretization Methods and Performances”, Proceesings of the International Conference on Electrical Engineering and Informatics Institut Teknologi Bandung, pp. 535 - 537, 2007.
18. K.I. Moin, Q.B. Ahmed, “Use of Data Mining in Banking”, International Journal of Engineering Research and Application, Vol. 2, Issue. 2, pp.738 - 742, 2013
19. D.H. Shih, H.S. Chiang, C.D. Yen, “Classification methods in the detection of new malicious emails”, Information Sciences, Vol. 172, Issue. 1-2, pp. 241 - 261, 2005.
20. K.G. Srinivasa, K.R. Venugopal, L.M. Patnaik, “A self-adaptive migration model genetic algorithm for data mining applications”, Information Sciences, Vol. 177, Issue. 20, pp. 4295 - 4313, 2007.
21. C.T. Su, J.H. Hsu, “An extended chi2 algorithm for discretization of real value attributes”, IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 3, pp. 437 - 441, 2005.
22. F. Tay, L. Shen, “A modified chi2 algorithm for discretization”, IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 3, pp. 666 - 670, 2002.
23. J.T. Tou and R.C. Gonzalez, “Pattern Recognition Principles. Addison-Wesley”, 1874.
24. C.-J. Tsai, C.-I. Lee, and W.-P. Yang. “A discretization algorithm based on class-attribute contingency coefficient”, Information Sciences, Vol. 178, pp. 714 - 731, 2008.
25. A.K.C. Wong, D.K.Y. Chiu, “Synthesizing statistical knowledge from incomplete mixed-mode data”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-9, Issue. 6, pp. 796 - 805, 1987.
26. Q. Wu, D.A. Bell, T.M. McGinnity, G. Prasad, G. Qi, X. Huang, “Improvement of decision accuracy using discretization of continuous attributes”, Proceedings of the Third International Conference on Fuzzy Systems and Knowledge Discovery, Vol. 4223, pp. 674 - 683, 2006.
27. R.R. Yager, “An extension of the naive Bayesian classifier”, Information Sciences, Vol. 176, Issue. 5, pp. 577 - 588, 2006.