落實 Bitcoin (Altcoins)/Blockchain 大數據分析於雲端平台
2018 年春季班「數據科學與大數據分析」
學期群組計畫需求說明書

胡毓忠

國立政治大學資訊科學系

hu@cs.nccu.edu.tw

2018 年 3 月 14/15 日

摘要

本學期「數據科學與大數據分析」學期群組計畫將以 Bitcoin (Altcoins) Core Blockchain 大數據資料集來進行區塊鏈 (Blockcahin) 的分析與學習模型建置和優化。參與此學期群組計畫的修課同學將能瞭解 Bitcoin (Altcoins)/Blockchain 的基礎知識、運作原理與技術，並進一步運用大數據分析的核心技術如資料集前處理、特徵值辨識、萃取、轉換與選取來探索和分析 Bitcoin (Altcoins)/Blockchain 虛擬貨幣付費交易結構圖來完成基礎圖論指標分析。此外各組需要更進一步選擇適當機器（統計）學習演算法經過學習模型建置與優化處理完整流程來達成所選取區塊鏈進行特定的大數據指標，如分群、分類、關連性、預測或因果影響力因子的學習與分析。

修課同學們將以 2-3 人分組來進行學期計畫系統設計、分析與實作，主要包含了有 Blockchain 資料集前處理、特徵質辨識、萃取，轉換、與選取並透過資料分析與模型建置來優化其分析結果完成有意義的結果解讀。最後各組分析與實作成果將以完整的口頭和書面報告當作為學期計畫成績評分依據。

學期分組計畫驗證分成兩階段：第一階段在期中考週完成本課程所提供 VMware 私有雲 Spark 分析引擎平台上的區塊鏈資料集的探索、處理與簡單淺層分析。第二階段則在期末考週各組選擇特定公有雲端平台如 Amazon AWS 或 Google GCP 的 Spark 分析引擎來完成 Bitcoin (Altcoins) Core Blockchain 的完整大數據區塊鏈資料集的深層與因果分析與影響力計算。

1 情境說明

Bitcoin (Altcoins)/Blockchain 的虛擬式密碼貨幣 (Cryptocurrency) 科技發展趨勢已經成為電腦科學與相關跨領域學門的主要研究與發展重點。2019 年 1 月 Bitcoin 正式上線運作後，經過不斷的成長其市值在 2017 年底已經到達 1700 億美金，區塊鏈每一天被確認交易量則大約有 37 萬 5 千比 (參考文獻 (d))。Bitcoin 開放式區塊鏈運作因為有其特殊性，透過分散式資料庫的存放與處理來確認其無法被修改區塊鏈的安全性與完整性。本學期大數據群組計畫將以 Bitcoin (or Altcoins) 的 Blockchain 大數據為主要分析標的。各組首先需要瞭解 Bitcoin (Altcoins) 其 Blockchain 的運作原理，以便能夠探索線性串連式區塊鏈來有效進行資料集內在既有特徵值的探索、前處理、與萃取並以簡單式淺層式分析 (參考文獻 (g))，如使用 MapReduce 高效能平行計算各種在特定時間範圍內的 Bitcoin (Altcoin) 區塊鏈的指標如總交易量，最大交易量，總交易與最大交易手續費，和交易數量分布圖等。

更進一步分析者能夠完成 Bitcoin(Altcoin) 虛擬貨幣交易圖論的深層分析，如特定時段內的付費交易連結圖來進行使用者交易如何分群並起發掘可能異常交易的完整拓普圖結構的產出，或透過 PageRank 指標計算來找出虛擬貨幣付費交易圖中其影響力最大的前幾筆交易 (參考文獻 (f)(l))。最後可以考慮結合外在資料源如 Bitcoin 和美金（歐元）的匯率交換價格特徵值來分析與預測 Bitcoin 的價格變動 (參考文獻 (h))，或分析有那些內在與外在因子會影響或干擾 Bitcoin (Altcoins) 價格變化以及可能產生具體正（負）面影響力 (參考文獻 (c)(j)) ；或整合外在 Bitcoin (Altcoin) 社群媒體討論群使用者資訊來分析 Bitcoin 付費交易所產生的交易型態分類，以及可能的洗錢交易或其它犯罪行為偵測與判斷的二元分類和預測等。

本學期群組計畫需要完成下列三項目標：

1. 透過 MapReduce 分散式高效能運算處理來找出區塊鏈付費交易資料集的基礎淺層分析相關指標。

2. 選擇適當機器學習演算法如 linear regression, logistic regression, decision trees, regression trees, random forests, deep learning, etc 完成對於 Bitcoin Core Blockchain 完整資料集的模型訓練與測試流程。各組必須運用分析模型選擇與優化參數與超級參數的技巧如 Cross-Validation (CV), Regularization 並且參考可能最佳檢驗值如 minimum mean-square-error (MMSE), AIC, BIC, adjusted R^2, ROC/AUC 曲線與面積來找出最佳分類或價格預測模型 (參考文獻 (e)(o))。

3. 建議加分項目：使用 R Bayesian Structural Time Series (BSTS) *CausalImpact* 函式庫套件找出適當的干擾因子來進行 Bitcoin (Altcoin) Blockchain 價格或其它預測相依變數的影響力分析，並且計算出其具體反事實 (Counterfactual) 的數值 (參考文獻 (c)(j))。

2 需求與規格的分項配分

學期分組實作計畫佔本課程總成績的 50%。詳細計畫需求規格配分說明如下：

1. 各組的大數據分析系統實作分成兩階段：
   (a) 期中考週當日完成第一階段本課程所提供 VMware 私有雲 Spark 引擎上使用 MapReduce 技術進行 Bitcoin(Altcoin) Blockchain 的資料集基礎淺層分析。
   (b) 期末考週當日完成雲端版 (Amazon AWS 或 Google GCP)Spark 引擎上的完整 Bitcoin (Altcoin) Blockchain 大數據分群、分類或預測的深層分析。
   (c) 如果可能進一步對於上述深層分析指標完成特定內在或外在干擾因子對於其分群，分類或預測的反事實分析與影響力計算。
2. 強烈建議各組透過 Piazza/GitHub 合作平台來進行大數據分析的開發與實作，並提供完整歷史紀錄檔在學期計畫驗收與成果報告時協助老師對於各組員對於學期計畫的個別貢獻度評量。
3. 學期分組計畫的評分 (50%) 的分配如下：
   (a) 上述分項 (1)(a) 占 15% 並在期中考當日以口頭與簡式書面報告來進行檢視與評量。
   (b) 上述分項 (1)(b) 占 30% 並在期末考當日以口頭報告與後續完整 10 頁書面報告來進行檢視與評量。
   (c) 各組提供完整的 Piazza/GitHub 平台開發與實作記錄並且經過檢視分組計畫成績占 5%。
   (d) 上述分項 (1)(c) 為加分題，視實做結果可加分 5-10 分。
4. 本學期群組計畫的第一階段檢驗與報告日為期中考當天，即一般生為 2018/04/25 (星期三) 而碩專生為 2018/04/26(星期四)。第二階段檢驗與報告日為期末考當天，即一般生為 2018/06/27 (星期三) 而碩專班生為 2018/06/28 (星期四)。學期計畫 10(+-) 頁完整書面報告繳交截止日則另行公布。
5. 本需求說明書如有未盡事宜之處將在課堂中補充說明。

3 參考文獻與網址

1. 參考文獻：

(a) Bartoletti, M., et al., A general framework for blockchain analytics, *arXiv:1707.01021v2*, 6 Nov 2017.

(b) Bonnneau, J., et al., SoK: Research perspectives and challenges for bitcoin and cryptocurrencies, *2015 IEEE Symposium on Security and Privacy*, May 2015.

(c) Brodersen, H. K., et al., Inferring causal impact using Bayesian structural time-series models, *The Annals of Statistics*, 9(1), 2015, pp. 247-274.

(d) Conti, M., et al., A survey on security and privacy issues of bitcoin, *arXiv:1706.00916v3*, Dec., 2017.

(e) Fawcett, T., An introduction to ROC analysis, *Pattern Recognition Letters*, 27, 2006.

(f) Fleder, M., et al., Bitcoin transaction graph analysis, *arXiv:1502.01657v1*, 5 Feb. 2015.

(g) Kalodner, H., et al., Blocksci: design and applications of a blockchain analysis platform, *arXiv:1709.02489v1*, 8 Sep. 2017.

(h) Madan, I., Saluja, S., and Zhao, A., Automated bitcoin trading via machine learning algorithms, 2014.

(i) Narayanaa, A. and Clark, J., Bitcoin's academic pedigree, *Communication of the ACM*, Dec., 2017, pp. 36-45.

(j) Poyser, O., Exploring the determinants of bitcoin's price: an application of Bayesian Structural Time Series, *arXiv:1706.01437v1*, 5 June, 2017.

(k) Reid, F. and Harrigan, M., An analysis of anonymity in the bitcoin system, arXiv:1107.43524v2, 7 May, 2012.

(l) Ron, D. and Shamir, A., Quantitative analysis of the full bitcoin transaction graph, *Financial Cryptography and Data Security (FC 2013)*, pp. 6-24, 2013.

(m) Rubin J., BTCSpark: scalable analysis of the bitcoin blockchain using Spark, http://www.mit.edu/ jlrubin/projects/btcspark/

(n) Tschorsch, F. and Scheuermann, B., Bitcoin and beyond: a technical survey on decentralized digital currencies, *IEEE Communications Survey & Tutorials*, 18(3), 2016, pp. 2084-2123.

(o) Zheng, A., *Evaluating machine learning models: a beginner's guide to key concepts and pitfalls*, O'Reilly Media, Sep., 2015.

2. 參考網址：

(a) Bitcoin Magazine: https://bitcoinmagazine.com/

(b) Hadoopcryptoledger Wiki
https://github.com/ZuInnoTe/hadoopcryptoledger/wiki

(c) Fetching Bitcoin Core Blockchain Data:
https://github.com/ZuInnoTe/hadoopcryptoledger/wiki/Fetching-Bitcoin-Core-Blockchain-data

(d) The Blockchain Meets Big Data and Realtime Analysis
https://bitcoinmagazine.com/articles/blockchain-meets-big-data-realtime-analysis-1435183048/

(e) BitcoinBlock Explorer - Blochchain: https://blockchain.info/

(f) Coinalytcs/Skry: https://www.ibm.com/us-en/marketplace/7436
https://home.scry.info