【ICDE2014勉強会】

Session 1: Clustering

担当：孟 茜（同志社大学大学院）

- Incremental Cluster Evolution Tracking from Highly Dynamic Network Data
  - Pei Lee , Laks V.S. Lakshmanan, Evangelos E. Milios
- Finding Common Ground among Experts' Opinions on Data Clustering: with Applications in Malware Analysis
  - Guanhua Yan
- Towards Effective and Efficient Mining of Arbitrary Shaped Clusters
  - Hao Huang, Yunjun Gao,Kevin Chiew,Lei Chen, Qinming He

Incremental Cluster Evolution Tracking from Highly Dynamic Network Data

- 動的ネットワーク環境
  - ノイズが多い，scaleが多い，進化が早い
  - 本研究はhighly dynamicネットワーク環境ではクラスタの進化を追跡することに着目する
- 先行研究と問題点
  - 先行研究はクラスタの進化追跡することには，node-to-nodeの方法で，クラスタ維持をしている．パフォーマンスが低い
  - 問題点
    - 動的ネットワーク環境は多次元で変化が早いので，増量計算は難題の一つ
    - 増量計算では，フォマットとクラスタ進化の追跡操作が実現しにくい
    - 膨大なアップデートを処理するのは難しい

提案手法1

- Step
- 1.Skeletal Graph（オリジナルのpost networkの簡潔な要約）
- Fading time window を利用し更新の監視、進化パターンの取得。原始進化の操作と代数計算でクラスタ化する
- Node-by-node をsubgraph-by-subgraphに拡張し，クラスタ進化追跡のパフォーマンスの向上

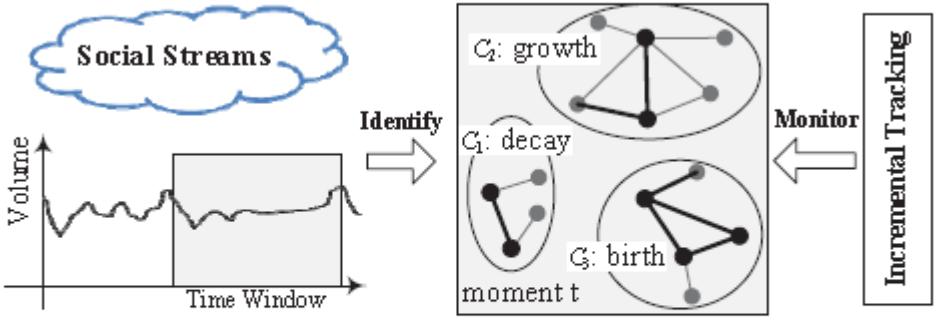
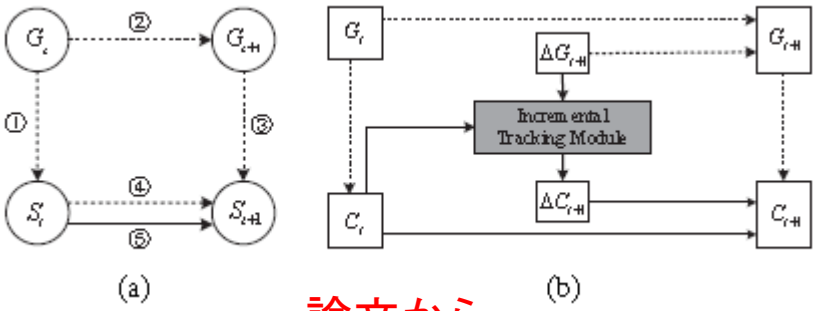


Fig. 1. Post network captures the correlation between posts in the time window at each moment, and evolves as time rolls on. The skeletal graph is shown in bold. From moment t to $t+1$, the incremental tracking framework will maintain clusters and monitor the evolution patterns on the fly.

論文から

提案手法2

- Skeletal Graphラスタの枠組
  - Post network構築
    - ソーシャルストリームの予備処理
      - 従来の方法はエンティティの情報を提供不足で，ここはポストPの表示は(L, τ, u)，p^Lはエンティティのリスト，p^τはスマップ，p^uはアーサ
    - Fading 類似度
      - ポストの作成時間が連接であれば，類似の可能性高い，内容だけではなく，時間も考慮すべき．指数関数を利用し，時間の流れを表現する．
  - Node最適化
  - Skeletal Graphの定義
- 増量計算アルゴリズム



論文から

貢献

- 高度動的ネット環境の増分計算フレームワークの提案
- 従来の処理は一つのノードずつになり，提案手法ではSkeletal graphを利用し，一気に増，削，進化を処理できる
- Bulk updatingに対して増量計算への2つの提案．増量クラスタ維持にICM，クラスタ進化追跡にeTrack
- Twitter streamsの実験で，提案が効果よく追跡全ての進化パターンin time

Finding Common Ground among Experts' Opinions on Data Clustering: with Applications in Malware Analysis

- 背景
  - クラスタリングは計算のコストが高くなり，リソースの要求が高くなっていく．クラスタの計算タスクをほかの専門家アプリに任すこともある
  - クラスタリングの計算方法や距離の評価基準の違いは同じデータセットで違う結果が出る．
  - クラスタするできる対象は同じクラスタの対象は研究されたので，対応策がある．それをヒントとし，本研究は違うクラスタリングアルゴリズムの共通点を研究する．

提案手法

- それぞれのクラスタリング方法は違う結果が出るので，データセットの全てのデータをグループイングではなく，共通の部分集合を求める
- 手順
  - クラスタが共通かどうかの条件を厳しく決める．それぞれの専門家クラスタリング方法を利用し，クラスタを生成し，全てのアルゴリズムでは矛盾がない集合は求めるクラスタ
  - 矛盾のないデータオブジェクトを探す
  - 3つの設計（1つgreedy solution，2つランダムアルゴリズム）を実行し，矛盾がある最大独立集合を探す
  - k-partite graphを利用し，全ての設計をコネクトする

実験

- ネット上には，悪意ソフトが多数存在，不正解析の研究は重要になった
  - 不正解析に関する研究の1つは悪意ソフトのファミリー検出と進化計算
  - 5つのAVソフトを利用し，悪意ソフトのファミリーネームを検出した.
  - AV software検出結果では，総数が448,790件の悪意ソフトの実体を検出したので，original clustering matrix Cは448,790列がある．clustering matrix Dに143,101件として圧縮した．これは共通の悪意ソフトをクラスタリングする結果.

AV Software	Detection result	Family name
McAfee	Vundo.gen.m	Vundo
NOD32	a variant of Win32/Adware.Virtumonde.NBG	Virtumonde
Kaspersky	Trojan.Win32.Monderb.gen	Monderb
Microsoft	Trojan:Win32/Vundo.BY	Vundo
Symantec	Packed.Generic.180	GENERIC

論文P23

貢献

- データマイニングとデザインの時，専門家意見の共通点を計算する研究の1つとなる.
- 矛盾なしに専門家共通のクラスタリング方法を提示できる
- クラスタリングを行う時，公平的に一致性と質を保つ提案である.

Towards Effective and Efficient Mining of Arbitrary Shaped Clusters

- 背景
  - クラスタリングはデータの類似度に基づき，データをグループに分ける．どんなクラスタリング方法は最適かはデータセットに依存
  - 任意形クラスタのマイニングは広く応用された(空間データマイニング，画像分割，生物医学など)，それをマイニングするクラスタリング方法が少ない
- 問題点
  - The spectral, graph-based と density-basedクラスタリングはパターン分析に要する計算のコストが高い，膨大なデータセットでは軸限定の対応策を取った
  - サイズ収縮のクラスタリング方法は収縮過度や収縮不足の問題がある

提案手法1

- CLASP (Clustering　aLgorithm for Arbitrary ShaPed clusters) algorithm ではthree-phaseを提案した
  - クラスタ形を維持するうえで自動的にクラスタのサイズ削減
  - データセットから最もサポートできるエリアを抽出する
  - K-meansでそのサポートエリアを小さいグループに分解する
  - グループ中心が分散により自動的に選択される
  - その中心を元のデータセットの代表的なデータとしてみなす

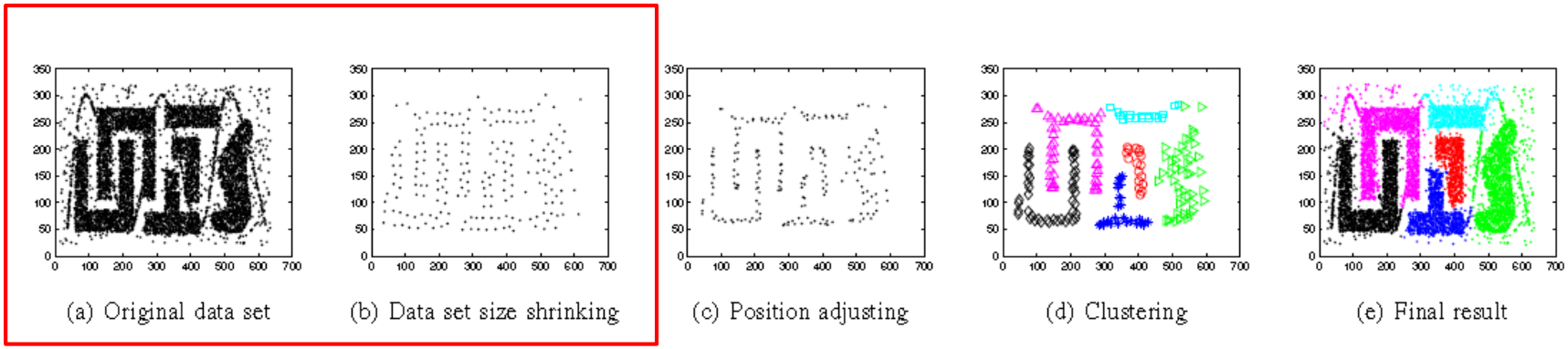


Fig. 1.　Example of the processing result in each phase of CLASP algorithm

論文P29

提案手法2

- クラスタの構築をはっきり，明確化にする．クラスタの精度を向上させる．

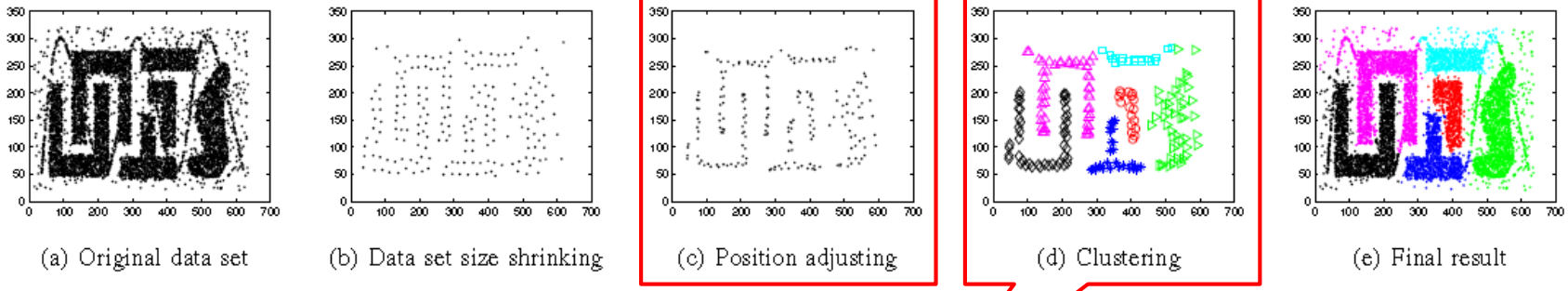


Fig. 1. Example of the processing result in each phase of CLASP algorithm

論文P29

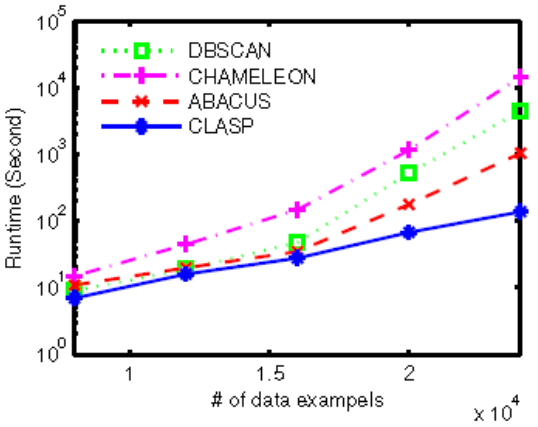
- 任意型クラスタをマイニングする．
  - グループの中心をMutual k-Nearest Neighborsアルゴリズムで距離の近いサブクラスタを結合する．その結果はサブクラスタがそれぞれの領域に任意型で分散していく

評価実験

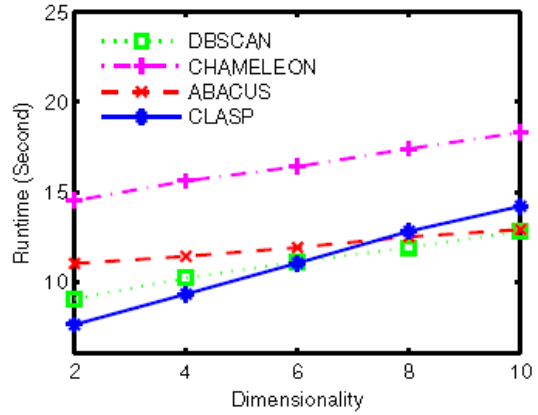
- 任意型のクラスタリングを適任できる

- 比較手法
  - density-based ：DBSCAN
  - canonical graph-based : CHAMELEON
  - state-of-the-art　shrinking-base：ABACUS

- Run-timeは提案手法が最も早い
- Seed（グループ中心）の初期値が合理的
- 効率の評価基準では最もよりパフォーマンス



(a) Scalability to number of data examples ($d =$



(b) Scalability to dimensionality ($n =$ 8000)

論文P37

Fig. 6. Algorithms' scalability study

貢献

- 提案したCLASPはクラスタリングの計算削減ができた
- 任意型のクラスタの形情報を維持できた
- 有用データサンプルの表現が明確化ではっきりしたクラスタリング実現できた

- 考察
- クラスタリングをする時，そのクラスタの大枠の情報を抽出し，クラスタ中心との距離を利用し，クラスタを計算方法は，クラスタの型を維持できる.