

類似ネットワークによる市場データの分析 Analysis of Market Data based on Similarity Networks

田中 雄大[†]
Yudai Tanaka

岩崎 清斗[‡]
Kiyoto Iwasaki

大久保 誠也[†]
Seiya Okubo

齊藤 和巳[†]
Kazumi Saito

1. はじめに

近年、Web上には膨大な量のマルチメディアデータが蓄積されており、これらデータ間の関係や特性を把握することは一層重要になっている。画像や音声、映像などに代表されるマルチメディアデータの多くは高次元ベクトルで表現することができる。この高次元ベクトルで定義されたオブジェクト集合を低次元ベクトルに埋め込むことは、データの隠れた構造やオブジェクト間の関係を視覚的に把握するために重要である。

本稿では、農産物(野菜)の品目別取扱実績の金額、数量、平均価格の時系列データをオブジェクトとし、類似ネットワークを構築して可視化する分析方法論について探求する。類似ネットワークの構成法としては、オブジェクト毎に最類似の k -オブジェクトに対しリンクを付与する方法(k -近傍ネットワーク) [1]、及び、任意のオブジェクトペアに対し他に両者の類似オブジェクトが存在しない場合にリンク付与する方法(相対近傍ネットワーク) [2]を採用する。15年間の月別市場データを用いた評価実験により、時系列データの種類とネットワーク構成法の違いによる特徴などを評価する。

2. 分析法

各オブジェクトは T -次元ベクトル $\mathbf{x}_n = (x_{n,1}, \dots, x_{n,T})$ で表されるとする。以下では、オブジェクト集合 $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ と、各オブジェクトペアに対する類似度 $s(\mathbf{x}_n, \mathbf{x}_m)$ が与えられたとき、 k -近傍法 [1]と相対近傍法 [2]のそれぞれで類似ネットワークを構築する方法について説明する。また本稿では、類似度として、機械学習やテキストマイニングの分野 [1]で標準的に採用されるコサイン類似度を用いる。ここで、コサイン類似度は次式で表される。

$$s(\mathbf{x}_n, \mathbf{x}_m) = \frac{\sum_{t=1}^T \mathbf{x}_{n,t} \mathbf{x}_{m,t}}{\sqrt{\sum_{t=1}^T \mathbf{x}_{n,t}^2 \sum_{t=1}^T \mathbf{x}_{m,t}^2}}. \quad (1)$$

各オブジェクトは類似ネットワークのノードに対応するため、以下では、オブジェクトをノードとも呼ぶ。

k -近傍法(以後 k -NN法と呼ぶ)は、パターン認識の分野でしばしば用いられる、単純な機械学習アルゴリズムである。詳細な手順を以下に示す。

1. $k = 1$ に初期化する
2. 各ノードは、自身との類似度が k 位までのノードとリンクを結ぶ
3. 全ノードが連結ならば2.の結果をネットワークとして出力する。さもなければ、 $k \leftarrow k + 1$ とし、2.に戻る。

[†]静岡県立大学 経営情報学部

[‡]静岡県工業技術研究所 電子科

k -NN法により構築したネットワークを k -NNネットワークと呼ぶ。

相対近傍法(以後RNG法と呼ぶ)は、任意のノードペアに対し、相対的に両ノードと類似するノードが存在しない場合にリンクを付与する。いま、任意のノード間の距離としてコサイン類似度から誘導されるユークリッド距離 $d(\mathbf{x}_n, \mathbf{x}_m) = \sqrt{1 - s(\mathbf{x}_n, \mathbf{x}_m)}$ を導入する。詳細な手順を以下に示す。

1. 任意のノードペア $\{\mathbf{x}_n, \mathbf{x}_m\}$ に対し、ノード間の距離 $r = d(\mathbf{x}_n, \mathbf{x}_m)$ を半径とする円を、ノード $\mathbf{x}_n, \mathbf{x}_m$ のそれぞれを中心にして描く(その重なった部分をLuneと呼ぶ)
2. Luneにその他のノードが存在しない場合に、ノードペア $\{\mathbf{x}_n, \mathbf{x}_m\}$ に対しリンクを張る
3. 得られた結果をネットワークとして出力する

すなわち、2つのオブジェクト \mathbf{x}_n と \mathbf{x}_m が相対近傍であるとは、次式が成立するときであり、このときリンクが生成される。

$$d(\mathbf{x}_n, \mathbf{x}_m) \leq \max_{k \notin \{n, m\}} \{d(\mathbf{x}_n, \mathbf{x}_k), d(\mathbf{x}_m, \mathbf{x}_k)\} \quad (2)$$

RNG法により構築したネットワークをRNGネットワークと呼ぶ。RNGネットワークでは全ノードの連結が保証される。

3. 実験による評価

本実験では、2002年1月から2016年12月までの期間で、品目別取扱実績(野菜)の金額、数量、平均価格の月別データを用いた[§]。また、データのうち、15年間のすべての月で欠損値のない品目に限定している。その結果、分析対象の野菜数は $N = 137$ となった。なお、この期間の月の総数は $T = 180$ となる。

図1ならびに図2に、数量、金額、平均価格のそれぞれのデータから構築した k -NNネットワークとRNGネットワークの可視化結果を示す。なお、本実験の可視化にはバネモデル法 [3]を用いた。

k -NNネットワークの可視化結果からは(a)金額がほかの野菜との結びつきが強くノード同士の結びつきの密度が高い。そして、金額や数量においては、複数のグループに分かれていることを捉えることができる。金額におけるグラフの右下に見えるのは、八つ頭等の12月に出荷が最盛期を迎える冬野菜のグループ、一方、左側に見えるのは、とうもろこし等の6月から7月に出荷が最盛期を迎える夏野菜のグループであった。また、数量のグラフにおける上に見えるのは、金額のグラフ

[§]<http://www.shijou-tokei.metro.tokyo.jp/>

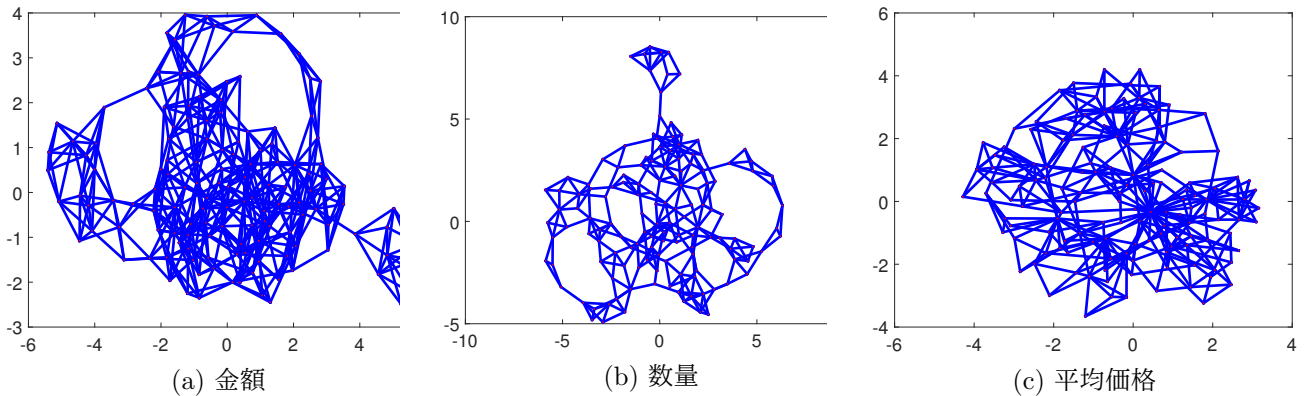
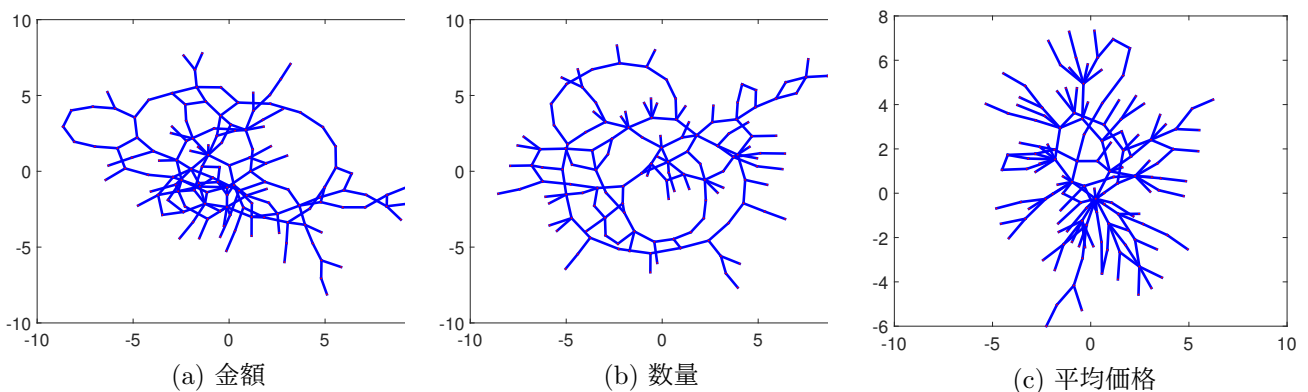
図1: k -NN ネットワークの可視化結果

図2: RNG ネットワークの可視化結果

における八つ頭と同様のグループであった。これらのグループは、それぞれの旬の季節にのみ大量の出荷がなされる野菜であり、 k -NN ネットワークから見えるグループ分けは適切であると言える。

RNG ネットワークの可視化結果からは (c) 平均価格が (a)(b) と比べて周囲に行くほど枝分かれしている。その一方で、全体的に疎な結合となり、グループのようなものを捉えることはできない。

数量に関するグラフ同士を比較すると、 k -NN ネットワークのグラフで上に見える八つ頭のグループは、RNG ネットワークではグラフの右上あたりに細長く伸びる枝として表現されている。このことから、RNG ネットワークでも特徴が出ていることがわかる。しかしながら、どこまでがグループに含まれるか否かの境界は曖昧であり、視覚的に捉えづらい可能性がある。

平均価格同士を比較すると、 k -NN ネットワークの場合は、全体が一塊として可視化されてしまっており、全体の特徴を捉えることが難しい。一方、RNG ネットワークの場合は、周囲に行くほど枝分かれした形状となり、同様の傾向があるオブジェクトは同じ方向に表現されている。

これらの結果から、今回利用した農産物(野菜)の品目別取扱実績に対しては、RNG ネットワークと k -NN ネットワークでは同様の趣旨を表現しているが、読み取りやすさに大きな差があることがわかった。

4. おわりに

本稿では、農産物(野菜)の品目別取扱実績の金額、数量、平均価格の時系列データをオブジェクトとし、RNG ネットワークと k -NN ネットワークを構築して可視化する分析方法論について探求した。15年間の月別市場データを用いた評価実験では、時系列データの種類とネットワーク構成法の違いによる特徴などを実証評価した。今後は、さらに多様なデータを対象に評価実験などを進める。

謝辞 本研究は、科学研究費補助金基盤研究(C)(No.15K00429)の助成を受けた。

参考文献

- [1] C. M. Bishop (2010), Pattern Recognition and Machine Learning (Information Science and Statistics), Springer.
- [2] J. K. Supowit (1983), The relative neighborhood graph, with an application to minimum spanning trees, Journal of the ACM (JACM) 30, pp. 428-448.
- [3] T. Kamada & S. Kawai (1989), An algorithm for drawing general undirected graph, Information Processing Letters, 32, pp. 7-15.