공간 데이터 마이닝을 활용한 은행고객분석
-강남·서초구를 중심으로-

최경희[1] · 황철수[2]

경희대학교 지리학과 대학원[1], 경희대학교 지리학과 조교수[2]

1. 서론

정보기술과 컴퓨터 기술의 급속한 성장에 따른 데이터의 양적 증가로 최근 다양한 분야에서 과학적이고 정교한 분석이 요구되고 있다. 특히 은행분야는 금융환경과 소비자의 태도변화로 오래 전부터 축적해온 고객에 대한 방대한 데이터를 효과적으로 분석하여 이를 통한 사업성 증대를 꾀하고 있다. 대부분의 은행의 고객 데이터 분석에서는 고객의 거래 정보나 인구통계 정보 분석이 대부분을 차지한다. 하지만 거래정보나 인구통계 정보 등은 데이터의 비공간적 속성으로 데이터가 지닌 중요한 속성인 공간적 속성은 간과한 채 모든 분석들이 이루어졌다. 또한 공간 데이터는 비공간 데이터와는 다른 특징을 지니고 있기 때문에 대용량의 공간 데이터를 분석하기 위해서는 기존 대용량의 비공간 데이터 분석에 사용되었던 데이터 마이닝 방법론이 아니 공간 데이터에 적절한 방법론 사용해야한다.

본 연구는 대용량의 고객 데이터 분석에서 다루지 않았던 공간적 속성을 중심으로 공간 데이터 마이닝 방법론을 적용하여 은행 고객의 공간적 패턴을 분석하는 것이 목적이다. 이를 위하여 고객 데이터의 공간적 요소를 확인하는 과정과 공간적 분포패턴을 확인하는 과정으로 나누어 분석을 실시하였다. 공간 데이터는 기존 데이터 마이닝에서 다루었던 비공간 데이터와는 다른 공간적 자기상관성과 공간적 이질성을 지니고 있으므로 은행 고객의 정확한 위치를 기반으로 하는 점 사상의 데이터와 동 단위를 기본으로 한 면 사상의 데이터로 분류, 공간통계기법을 통하여 공간 데이터의 특수성을 설명하였다. 이러한 공간적 요소를 바탕으로 공간 데이터 마이닝의 방법론 중 하나인 TIN을 이용한 계층적 군집 분석 방법을 이용하여 은행 고객의 군집 패턴을 찾아내고 특징을 분석하였다. 그후 주 은행의 위치와 타 은행과의 접근성이라는 변수를 이용하여 은행 고객의 공간적 분포패턴과 지리 사상들과의 연관성을 분석하였다.

2. 공간데이터 마이닝 (Spatial Data Mining)

공간 데이터 마이닝은 공간 데이터베이스 내에서 이전에 알 수 없었던, 잠재되어 있고 흥미로운 정보와 공간적 상관관계, 다양한 공간적 패턴을 찾아내는 과정으로 공간 데이터는 기존의 데이터 마이닝에서 사용하는 비공간 데이터와는 다른 특징인 공간적 자기상관성과 공간 이질성[1]이 존재한다. 따라서 기존의 통계적 방법론에 이론적 근거를 둔 데이터 마이닝 방법론들은 공간 데이터에 적합하지 않아 공간 데이터 마이닝 방밥론이 대두된 것이다.

지금까지 연구된 공간 데이터 마이닝 기법들은 크게 일반화, 공간 군집, 공간 연관규칙분석, 공간 회귀(spatial regression)로 분류할 수 있다. 일반화, 공간 군집, 공간 연관성 분석은 설명적 공간 데이터 마이닝, 공간 회귀 기법은 예측적 공간 데이터 마이닝에 속한다

1) 공간 이질성(spatial heterogeneity)이란 서로 다른 입지적 특성으로 공간적 또는 지역 차이가 발생하는 현상을 의미한다.

3. 공간적 요소 확인

이 논문은 2001년 1월부터 3월까지 A은행의 강남·서초구의 한 지점을 이용한 개인 고객 데이터와 다양한 GIS레이어(layer)를 기본 데이터로 하여 분석을 수행하였다.

기본적 분석인 인구통계적 분석을 실시한 결과 30대, 40대 연령층은 전체 고객의 50%이상, 계좌수에서도 50%이상을 차지하고 있었다. 그러나 평균 잔액의 경우 전체 고객의 25%(30대 7.5%, 40대 18.3%) 정도에 그치고 있었다. 30대, 40대 대신 전체 고객의 25%에 미치지 못하는 50대, 60대, 70대의 고객이 전체 평균잔액의 약 70%를 차지하는 것으로 나타났다. (50대 28.2%, 60대 19.8%, 70대 20.9%) 20대 고객들은 계좌수에서는 20%(1062개)를 차지하고 있었으나 평균잔액비율은 2.5%만을 나타내고 있었다.

1) 주소정보별 은행고객의 분포 패턴

연구 지역 내의 점 사상이 규칙적인 패턴을 나타내는지, 아니면 임의적으로 분포되어 있는지를 파악하고 부수적으로 만약 규칙적인 패턴이 나타나면 어떤 공간적 규모 내에서 현상이 발생하는지, 특정 공간 집합체나 공간 군집과 다른 요소들과 근접성이 어떠한 관련이 있는지를 분석하기 위하여 점 패턴 분석은(Spatial point pattern)을 실시하였다. 전역적 점패턴분석을 실시한 결과 A은행의 은행 고객의 위치는 북동에서 남서 방향을 띠고 있으며 중심점은 서초 2동, 은행 지점과 멀지 않은 곳에 위치함을알 수 있었다. 또한 최근린 거리 분석(Nearest Neighbor Distance)과 K함수(K-function)을 이용하여 은행 고객의 분포가 공간적으로 군집함을 확인하였다

2) 면 패턴 분석

공간적 자기상관 지수는 지도상에 나타나는 공간적 패턴을 일반화하고 정량화시켜 공간 현상의 복잡한 과정을 이해할 수 있을 뿐 아니라 공간 상호작용 모형(spatial interaction model)과 같은 공간 과정을 개념화하고 이를 통하여 예측모델을 만들 수 있으며 이러한 모델을 통계적으로 진단하는데 유용한 방법으로 공간적 자기상관 지수를 이용하여 은행고객의 면패턴 분석을 실시하였다. 분석결과, Moran' I 값은 3.172(p=0.015)으로 강남·서초구의 은행고객의 분포는 매우 높은 공간적 자기상관이 존재하고 있음을 확인하였다. 또한 은행고객의 공간적 분포의 Local Moran을 확인한 결과 서초1, 서초2, 서초3동, 서초4동과 양재1동으로 공간 군집이 발생하는 지역이고, 개포3동, 대치동, 일원본동과 일원2동은 공간적으로 관련이 적은 지역으로 나타났다.

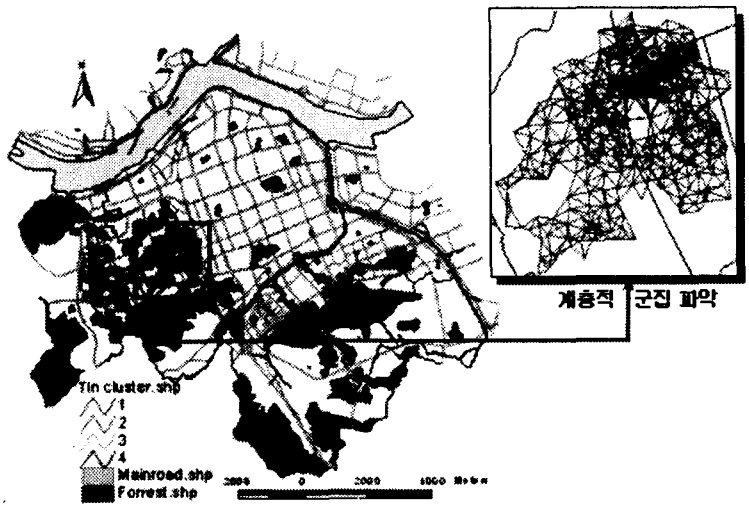
3. 공간적 특성 분석

1) TIN을 이용한 계층 군집 분석(hierarchical Clustering with TIN)

군집 분석은 서로 연관성이 높은 데이터들을 하나의 그룹으로 분류하는 작업으로 많은 수의 자료를 몇 개의 군집으로 분류하여 자료들에 대한 유용한 정보를 제공하는 방법으로 공간 군집 분석은 공간 데이터들의 속성이나 지리적 특성에 기반하여 서로 연관성이 있는 객체들을 그룹화하는 방법이다. 본 연구에서는 다양한 공간 군집 방법 중 TIN을 이용한 계층 군집 분석(Hierarchical Clustering with Triangular Irregular Network)을 이용하였다. TIN을 이용한 계층 군집 분석 방법은 유클리드 거리(Euclidean Distance)로 공간 근접성을 정의했던 기존의 방법과 달리 거리 모델의 단점을 극복한 모델로 우선 공간 인접성을 정의한 후 보로노이 다이어그램(Voronoi Diagram, VD)

을 통하여 공간이웃을 정의한다.[2] 군집분석에 사용되는 다양한 값들은 TIN을 통하여 얻어지며 공간 군집의 수를 정해주지 않아도 되기 때문에 연구자가 임의적으로 정의해야 하는 변수(user argument)가 필요 없다. 고밀도(high density) 군집 뿐 아니라 저밀도(low density) 군집도 분석 가능하며 다양한 모양의 군집을 모두 구할 수 있다. 또한 여러 종류의 일반화된 보로노이 다이어그램에 서로 다른 가중치를 적용하여 좀더 효과적인 군집을 만들 수 있다. 또한 공간 근접성을 정확히 모델링하여 전역적, 국지적 효과에 대한 정확한 정보를 얻을 수 있다.

은행고객의 공간적 군집성을 분석하기 위하여 은행 고객의 점 사상을 기반으로 TIN을 구성하고 적합한 threshold 값를 찾아 그 값보다 긴 TIN의 edge들을 제거하여 첫 번째 군집의 경계를 정했다. 그 후 edge를 재설정하여 이례지점과 연결된 edge들과 군집을 이루지 않는 edge들을 제거하였다. 이때 threshold는AMOEBA 프로그램의 식[3]을 사용하였다.(Estivill-Castro, V. and I. Lee, 2000)



TIN을 이용한 계층적 군집

2) 연관성분석

은행 고객의 군집 패턴의 확인은 고객들의 공간적 분포의 원인을 제대로 설명해줄 수 없다. 따라서 지리적 요소가 고객의 은행 사용에 있어 중요한 요인이라는 가정 하에 지리적 요소와 은행 고객의 공간적 분포와의 연관성을 분석하여 고객의 공간적 분포 원인을 파악하였다. 본 연구에서는 은행과의 접근성과 타 은행과의 접근성이라는 두가지 변수와 은행 고객의 공간적 분포 패턴, 이에 따른 은행 상품 소비패턴과의 연관성을 살펴보았다. 은행과의 거리가 1000m 내에 전체 고객의 49%가 위치하고 이들의 평균 계좌액은 전체의 73%, 카드나 외환거래에서도 51%가 이루어지고 있었다. 강남구와 서초구를 구분하는 경부고속도로라는 지리 사상에 의하여 은행고객의 패턴이 상이하게 나타나고 있었다. 경부고속도로에 의하여 분리된 강남구의 은행 고객들은 접근성이 양호한 서초구의 은행 고객들에 비하여 평균 계좌액이나 최종 계좌액이 높았다. 타 은행과 고객과의 관계는 거리가 증가함에 따라 점차적으로 고객의 거래액이 증가하는 양의 상관관계를 나타내고 있었다.

---

2) 보로노이 다이어그램을 공유하고 있는 점들을 공간 이웃으로 정의한다.

3) *threshold value = Global Mean + T(p)*

4. 결론

본 연구에서는 은행고객의 공간적 특성을 파악하기 위하여 우선 공간적 요인의 유무를 검토하고 이후 은행고객의 공간적 특성을 확증하였다. 본 연구의 분석결과들을 통해 다음과 같은 구체적 성과를 얻을 수 있었다.

기존 은행 데이터 분석에서 사용되지 않았던 공간적 요인을 주요한 분석 요인으로 선택하여 이에 적절한 방법론을 적용하였다. 기존에는 고객 분석에서는 적용되지 않았던 공간 통계 분석을 적용하여 데이터 내 공간적 요인의 유무를 확인하고 분포 패턴을 파악하였으며 TIN을 이용한 계층적 군집분석을 통하여 은행 고객의 공간 군집을 시각화, 군집간의 특징을 분석하였다. 또한 분석 결과 은행 고객의 주거 형태와 그 지역의 경제적 특징, 연구지역의 주요 지리적 사상이 은행 고객의 공간적 분포, 소비패턴에 영향을 주었으며 은행과의 접근성, 타 은행과의 거리라는 변수가 여전히 은행 고객 분석에 있어 중요한 변수임을 확인하였다. 따라서 이러한 연구를 은행 마케팅의 중요한 부분인 은행 위치 선정이나 고객 분석, 대상 고객 선정에 있어서 중요한 자료로 사용할 수 있다.

참고문헌

Anselin, L. and S. Bao, 1997, Exploratory Spatial Data Analysis Linking SpaceStat and ArcView, in Fischer, M.M., and Getis, A. (eds.), Recent Developments in Spatial Analysis: Spatial Statistics, Behavioural Modelling, and Computational Intelligence, Berlin: Springer-Verlag, 35-59.

Bailey, C. T. and A. C. Gatrell, 1995, Interactive Spatial Data Analysis, Longman Scientific and Technical.

Eldershaw, C. and M. Hegland, 1997, Cluster Analysis using Triangulation, Computational Techniques and Applications: CTAC97, 201-208.

Estivill-Castro, V. and I. Lee, 2000, AMOEBA: Hierarchical Clustering Based on Spatial Proximity Using Delaunay Diagram, Proceedings of the 9th International Symposium on Spatial Data Handling, 10-12.

Estivill-Castro, V. and I. Lee, 2000, AUTOCLUST: Automatic clustering via boundary extraction for mining massive point-data sets, 5th International Conference on Geocomputation, 23-25.

Gold, C. M., 1991, Problems with Handling Spatial Data - The Voronoi approach, CISM Journal ACSGC, 45(1), 65-80.

Karypis, G., E. H. Han and V. Kumar, 1999, Chameleon : Hierarchical Clustering Using Dynamic Modeling, IEEE Computer, 32(8), 68-75.

Paul, L and C, Graham, 1995, GIS for Business and Service Planning, Geoinformation International.

Miller, J. Harvey and Jiawein Han, 2001, Geographic Data Mining and Knowledge Discovery, Taylor & Francis, London and New York.