Eötvös Loránd University
Faculty of Informatics
Ph.D. School of Information Science
Ph.D. Program of Information Systems

Thesis Book

Ágnes Vathy-Fogarassy

Novel Graph Based Clustering and Visualization Algorithms for Data Mining

Supervisor:

Attila Kiss

associate professor, Ph.D.

Budapest, 2007.

1. Introduction

The amount of the data stored in various information systems grows very fast. These data sets contain not only known information, but new knowledge as well. Data mining is one of the most effective methods for exploring useful information from large data sets. Clustering, as a special area of data mining is, one of the most commonly used methods for discovering the hidden structure of the considered data set. The main goal of clustering is to divide objects into well separated groups in a way that objects lying in the same group are more similar to each another than to objects in other groups.

In the literature several clustering and visualization methods can be found. However, due to the huge variety of problems and data sets, it is a difficult challenge to find a powerful method that is adequate for all problems. The diversity of clustering problems resulted in several algorithms that are based on various approaches. A graph, as a complex representation form stores information about the data objects themselves or about their representative elements, and it also provides information about the relations of the objects or their representatives. Thereby, graphs may be suitable starting points for clustering.

The visualization of the data set plays an important role in the knowledge discovery process. In practical data mining problems usually high-dimensional data is to be analyzed. Since humans have difficulty in comprehending high-dimensional data, it is very informative to map and visualize the hidden structure of the complex data set into a low-dimensional vector space. As a graph also stores information about the relations of the objects, visualization methods based on graph-theory enhance the process of visual data analysis by revealing the relations.

2. Goals and Applied Methods

The objective of the present thesis work is to develop novel graph based clustering and visualization methods that are able to eliminate the drawbacks of the well-known methods, so that they could process data sets better and could provide new application possibilities. For this purpose this dissertation utilizes ideas and methods from the following research areas: data mining, clustering, graph-theory, neural networks, data visualization, dimensionality reduction, fuzzy methods and topology learning.

3. New Scientific Results

Thesis 1. – Improvement of Jarvis-Patrick clustering

In this thesis I have proposed an improved version of the Jarvis-Patrick clustering. The Jarvis-Patrick clustering utilizes the nearest neighbor approach to cluster the objects. The main disadvantages of the Jarvis-Patrick clustering are that it utilizes a very rigid decision criterion to classify the objects, and this decision criterion is only confined to the *k* nearest neighbors. To solve these problems I have defined a new similarity measure based on the nearest neighbors of the objects. This similarity measure is not restricted to the direct neighbors, but it can also take into account objects that are further away. Furthermore, the suggested similarity measure fuzzifies the crisp decision criterion of the Jarvis-Patrick algorithm. The combination of the proposed similarity measure with hierarchical clustering methods provides an effective tool for exploring groups of data.

These new scientific results are published in [4].

Thesis 2. – New clustering algorithm based on minimal spanning tree and Gath-Geva clustering method

I have proposed a new cutting process for the minimal spanning tree based clustering methods and based on this cutting criterion I have proposed a new clustering algorithm based on the minimal spanning tree of the objects and the Gath-Geva clustering method. Graph based clustering algorithms find groups of objects by eliminating inconsistent edges of the graph representing the data set to be analyzed. The resulting subgraphs yield clusters. However, due to the huge variety of problems and data, it is a difficult challenge to identify the inconsistent edges of graphs. To solve this problem I have suggested a new cutting process of graphs, that iteratively finds the best partitions based on the measure of the fuzzy hyper volume of clusters. Based on this cutting criterion I have also suggested a new graph based clustering algorithm, which is an effective combination of the minimal spanning tree based clustering and the partitional Gath-Geva algorithm. The suggested algorithm is able to solve the typical problem of the graph based clustering algorithms (chaining effect) and it also solves the initialization problem of the Gath-Geva algorithm. The resulting clusters of the proposed algorithm are easily interpretable with a compact parametric description.

These new scientific results are published in [4,6].

Thesis 3. – New topology based visualization method

In this thesis I have proposed a new topology based visualization method. As in practical data mining problems high-dimensional data has to be analyzed, it is very informative to map and visualize the hidden structure of the complex data set in a low-dimensional space. However, the data set to be analyzed often includes lower-manifolds that are nonlinearly embedded in a higher-dimensional vector space. In this thesis I have suggested a new graph based visualization method to unfold the real structure of data. The proposed method combines the main benefits of the topology representing networks and the multidimensional scaling. The suggested method is able to unfold and visualize the nonlinearly embedded manifolds in a low-dimensional vector space. In this thesis it has been shown, that the proposed method demonstrates good mapping qualities both in distance and neighborhood preservation of the objects, and it outperforms other well-known topology based visualization methods.

The publication of these new scientific results are in progress [1,2,3].

4. Utilization of Results

The present thesis contains new graph based clustering and visualization methods. The developed methods can be applied in many scientific areas due to the generality of the problems and their solutions. For example, in medical sciences graphs may be applied in the analysis of patients' paths or in biochemical researches to identify the domain structure of proteins. Naturally, in the area of information science we can also find several application possibilities. For example, the pages and hyperlinks of the World-Wide Web may be viewed as nodes and edges in a graph. The analysis of the connections of the Web is a popular research area. Furthermore, social network of people is a current topic. On the internet there are several community sites (e.g. iWiW) that help drawing the social network of people. The analysis of such social networks is also an interesting topic. Additionally, as the proposed methods work based on general principles, they can be integrated in arbitrary data mining program packages.

5. Future Work

The aim of this research was to develop novel graph based clustering and visualization techniques. These new scientific results raise some other interesting questions. The intrinsic dimensionality of a data set is usually defined as the minimal number of parameters or latent variables required to describe the data. As it was mentioned previously, the error values referring to the distance preservation capabilities of the mappings are in connection with the local intrinsic dimension of the manifold. It calls for further research to develop a new method that can estimate the local intrinsic dimension of a data set using these error values. Furthermore, the estimation of the dimensionality of fractals by these mapping qualities is also an interesting question. It would also be an interesting work to develop a mapping process which is able to preserve the dimensionality of the fractals.

Publications

The Author's Publications Related to the Dissertation

[1] **A. Vathy-Fogarassy**, J. Abonyi: Local and Global Mappings of Topology Representing Networks. *Information Sciences*, Elsevier (submitted)

[2] **A. Vathy-Fogarassy**, A. Werner-Stark, B. Gal, J. Abonyi: Visualization of Topology Representing Networks. In *Proceedings of IDEAL 2007, Lecture Notes in Computer Science series*, Springer Verlag (accepted)

[3] **A. Vathy-Fogarassy**, A. Kiss, J. Abonyi: Topology Representing Network Map – A new Tool for Visualization of High-Dimensional Data. *LNCS Transactions on Computational Science*, Springer-Verlag, 2007. (accepted)

[4] **A. Vathy-Fogarassy**, A. Kiss, J. Abonyi: Improvement of Jarvis-Patrick Clustering Based on Fuzzy Similarity. *Lecture Notes in Computer Science: Applications of Fuzzy Sets Theory*, Volume 4578/2007, ISSN: 0302-9743, pp. 195–202, 2007.

[5] **A. Vathy-Fogarassy**, A. Kiss, J. Abonyi: Hybrid Minimal Spanning Tree and Mixture of Gaussians based Clustering Algorithm. *Lecture Notes in Computer Science: Foundations of Information and Knowledge Systems*, Volume 3861/2006, ISSN: 0302-9743, pp. 313–330, 2006.

[6] **Á. Vathy-Fogarassy**, B. Feil, J. Abonyi: Minimal Spanning Tree based Fuzzy Clustering. *Transactions on Enformatika, Systems Sciences and Engineering*, Volume 8, ISSN: 1305-5313, pp. 7–12, 2005.

Other Publications

[7] **Dr. Fogarassyné Vathy Ágnes**: Csoportosítás (klaszterezés) – Chapter 4. (pp. 131–184.) in *Adatbányászat a hatékonyság eszköze – Gyakorlati útmutató kezdőknek és haladóknak,* Editor: Dr. Abonyi János, ISBN: 9636183422, ComputerBooks Kiadói Kft, 2006.

[8] Starkné W. Á., **Fogarassyné Vathy Á.**, Csoma Á.: Szakértő szoftverágens a diszlexia lehetőségének megállapítására. *Acta Agraria Kaposváriensis,* Vol. 10, No 1, 65–82. o., ISSN 1418-1789, 2006.

[9] Vassányi I., Rovnyai J., **Fogarassyné Vathy Á.**, Tobak T.: Adatbányászati alkalmazások az egészségügyben. *Informatika és Menedzsment az Egészségügyben,* 2006/5 szám, 49–53. o., 2006.

[10] **Á. Vathy-Fogarassy**, G. Balázs, T. Tobak, I. Vassányi: Intelligent Data Analysis Center: A Client/Server Mining Model over the Internet. In *Proceedings of 1st ADBIS Workshop on Data Mining and Knowledge Discovery (ADMKD'2005),* pp. 57–65, 2005.

[11] **Dr. Fogarassyné Vathy Á.**, Dr. Vassányi I.: Adatbányászati technológiák az egészségügyben. *Informatika és Menedzsment az Egészségügyben,* 2005/3 szám, 46–50. o., 2005.

[12] **Fogarassy-Vathy Ágnes**: Application of Data Mining Methods in Health Care Databases. In *Proceedings of the 6th International Conference on Applied Informatics,* Vol. I., pp. 261–266, 2004.

[13] **Dr. Fogarassyné Vathy Á.**, Dr. Pataricza A.: Intelligens Adatelemző Központ létrehozása IKTA 142/2002. electrical publication of *XXIII. Centenáriumi Neumann Kollokvium elektronikus kiadványa*, 2004.

[14] **Dr. Fogarassyné Vathy Á.**, Dr. Fogarassy Gy.: Egészségügyi adatok előkészítése elemzések céljából. *Informatika és Menedzsment az Egészségügyben,* 2003/8 szám, 36–41. o., 2003.

[15] **Vathy Ágnes**: Adatbázisok biztonságának tervezési módszere és megvalósítási kérdései. electrical publication of *Informatika a felsőoktatásban 2002,* Debreceni Egyetem, 2002.

[16] **Vathy Ágnes**, Kiss Attila: Database Security – Access Rights from Design to Implementation. In *Proceedings of the 5th International Conference on Applied Informatics,* Vol. I., pp. 85–94, 2001.

[17] **Vathy Ágnes**, Timár Lajos: Az EER modell koncepcióinak érvényesülése egy relációs adatbázis-rendszerben. *Informatika a felsőoktatásban '99 Konferencia kiadvány,* I. kötet, 92–97. o., Debreceni Egyetemi Szövetség, 1999.

[18] Timár Lajos, **Vathy Ágnes**: A jó minőségű EER modell helye az adatbázis-tervezésben. *Informatika a felsőoktatásban '99 Konferencia kiadvány,* I. kötet, 86–91. o., Debreceni Egyetemi Szövetség, 1999.

[19] Timár L., **Vathy Á.**, Vígh K. Telekesi É., Tátrai J., Szigeti J. Kocsis T., Vass I.: *Építsünk könnyen és lassan adatmodellt!*, Veszprémi Egyetem és Műszertechnika Kft., 1996.

[20] **Vathy Ágnes**, Németh Krisztián: *Adatmodellezési feladatok I.,* Veszprémi Egyetemi Kiadó, 1996.