Advanced Technology and Social Media Influence on Research, Industry and Community

Reda Alhajj(✉)

Department of Computer Science, University of Calgary, Calgary, AB, Canada
alhajj@ucalgary.ca

Abstract. The rapid development in technology and social media has gradually shifted the focus in research, industry and community from traditional into dynamic environments where creativity and innovation dominate various aspects of the daily life. This facilitated the automated collection and storage of huge amount of data which is necessary for effective decision making. Indeed, the value of data is increasingly realized and there is a tremendous need for effective techniques to maintain and handle the collected data starting from storage to processing and analysis leading to knowledge discovery. This chapter will cite our accomplished works which focus on techniques and structures which could maximize the benefit from data beyond what is traditionally supported. In the listed published work, we emphasized data intensive domains which require developing and utilizing advance computational techniques for informative discoveries. We described some of our accomplishments, ongoing research and future research plans. The notion of big data has been addressed to show how it is possible to process incrementally available big data using limited computing resources. The benefit of various data mining and network modeling mechanisms for data analysis and prediction has been addressed with emphasize on some practical applications ranging from forums and reviews to social media as effective means for communication, sharing and discussion leading to collaborative decision making and shaping of future plans.

Keywords: Social media · Social networks · Data analysis · Big data
Frequent pattern mining · Clustering · Bioinformatics

1 Introduction

Data is a major resource for decision making. Its value and importance has never been ignored since the existence of mankind on earth. It has been collected, stored and maintained using a wide variety of affordable means ranging from primitive to advanced. Indeed, collecting, storing and maintaining data was a cumbersome task in the past, mainly prior to the development of various technologies that gradually helped humans in handling data. However, the recent development in technology rapidly influenced data collection, storage and maintenance. For instance, sensors are becoming

© IFIP International Federation for Information Processing 2018
Published by Springer International Publishing AG 2018. All Rights Reserved
A. Amine et al. (Eds.): CIIA 2018, IFIP AICT 522, pp. 1–9, 2018.
https://doi.org/10.1007/978-3-319-89743-1_1

popular in all aspects of the daily life; they have been installed in almost every indoor and outdoor equipment. They are widely available and equipped with wireless communication skills which allow them to feed huge amounts of data that should be captured, stored, cleaned, and processed for knowledge discovery as main ingredient of effective decision making.

In the past, humans used computing devices in a limited way. Database management systems were developed to facilitate flexibility in storing and retrieving data. Making sense of data was left to domain experts who are expected to retrieve and study data related to a specific problem in a way to draw some conclusions which may guide the decision-making process. Automating the knowledge discovery process was better realized towards the end of the 20^{th} century when various machine learning and data mining techniques were developed and put in practice to serve a variety of application domains including business, health, security, etc.

To cope with the new era, researchers, developers and practitioners realized the need to develop new techniques and technologies capable handling growing volumes of data captured incrementally from heterogeneous sources. In other words, growth in volume and types of data expected to be processed suddenly witnessed a boom. Social networks and social media platforms are gaining increased popularity and are generating tremendous amounts of data. Surveillance devices are available almost everywhere. Even traditional archives are digitized. Consequently, storage media and techniques which were previously accepted as sufficient are no more capable of handling new needs. For instance, hard drives of personal computers were only couple megabytes in capacity when they were initially manufactured with less than one megabyte of main memory. People were happily competing to get the honor of owning and using such devices. It may be impossible to image using same computing platform in the current era where gegabytes of storage are no more sufficient. In fact, computing resources have improved rapidly to partially meet human needs but will never be satisfactory. Therefore, researchers and developers are always seeking new technologies and techniques, and hence conducting and advancing research will continue to attract more attention and investment.

Explicitly speaking, data volume, characteristics and associated expectations may be described as a moving target. This necessitates the availability of enough room for storage and sophisticated techniques for processing. People will continue to collect more data as time passes, but they will never afford to increase their computing power to handle their data effectively. Thus, the need for algorithms and techniques that can depend only on limited computing resources to deal with various aspects of data from dynamicity to volume, among others. Along this direction, we contributed various techniques and algorithms that could successfully satisfy a variety of applications which require handling large volumes of dynamic and stream data. These techniques are described in our published papers listed in the references at the end of this paper. Scalability is the main aspect considered by our techniques, including frequent pattern mining, clustering, network analysis, finding repeating patterns in long sequences, etc.

2 Partial Mentioning of Our Achievements

Our completed and ongoing research addresses various aspects of data from definition to construction to manipulation and analysis leading to knowledge discovery for decision making. Our initial contributions focused on traditional aspects related to handling and manipulation of data which were popular during the last two decades of the 20th century. We then gradually moved onto more advanced techniques which we realized as necessities since 1990s. These techniques, include, network analysis, data mining and machine learning techniques which have tremendously and visibly served various applications. We also realized scalability as a serious need especially in the current era of big data. We developed advanced techniques and adapt them to various domains, including:

- Bioinformatics and Health informatics
- Data partitioning and allocation
- Homeland security and terror/criminal network analysis
- Financial data analysis: from stock market to FOREX to fraud detection
- Web/network data analysis: from structure to content to usage
- Social media analysis and opinion mining including spam detection
- Recommendation and customer behavior analysis
- Network representation is a powerful mechanism for modeling many-to-many relationships.

A network consists of a set of nodes corresponding to the entities in the application domain and a set of links representing certain types of relationships between the entities. On the other hand, data mining includes a set of powerful techniques for studying the relationships/connections between various objects. Further, data mining may be used in the network construction phase. To construct a network data may be first analyzed using techniques like frequent pattern mining or clustering. Once a network is constructed, it can be analyzed for knowledge discovery.

Most of the traditional approaches for frequent pattern mining assume unlimited main memory which is not realistic. Therefore, scalability is a major concern when it comes to practical applications where data streams dynamically and available in large volume. To tackle these problems, we developed a novel approach which satisfies the following:

- The ability to mine in a bounded amount of memory space that may vary based on task priority. Thus, it is possible to mine using common PC.
- Improve external data access and make the mining process more I/O conscious.
- Introduce a specialized mining task aware memory manager for both RAM and the external memory.

We build a tree structure namely, Frequent-Pattern (FP) tree, which summarizes the given data and allows for effective discovery of frequent patterns. Each branch represents at least one transaction. We build the tree from left to right and from top to bottom as shown in Fig. 1.

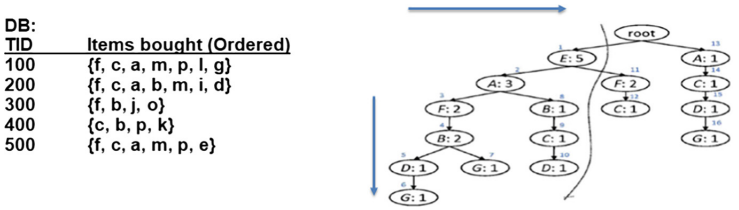

Fig. 1. Construction of FP-tree top-down and left-to-right

This way, we can store on the disk left side of the tree as it grows to the right. Therefore, our upper bound is the size of free disk space rather than the available memory.

To facilitate effective data investigation and analysis, we build our own tool, namely NetDriller which is capable of analyzing raw data to derive a network. Then various network analysis techniques could be applied on the network to identify actors which may reveal some important aspects related to the analyzed network, like most knowledgeable employee, most dangerous criminal, least performing student, best team to undertake next project, etc. The basics of NetDriller are summarized in Fig. 2.

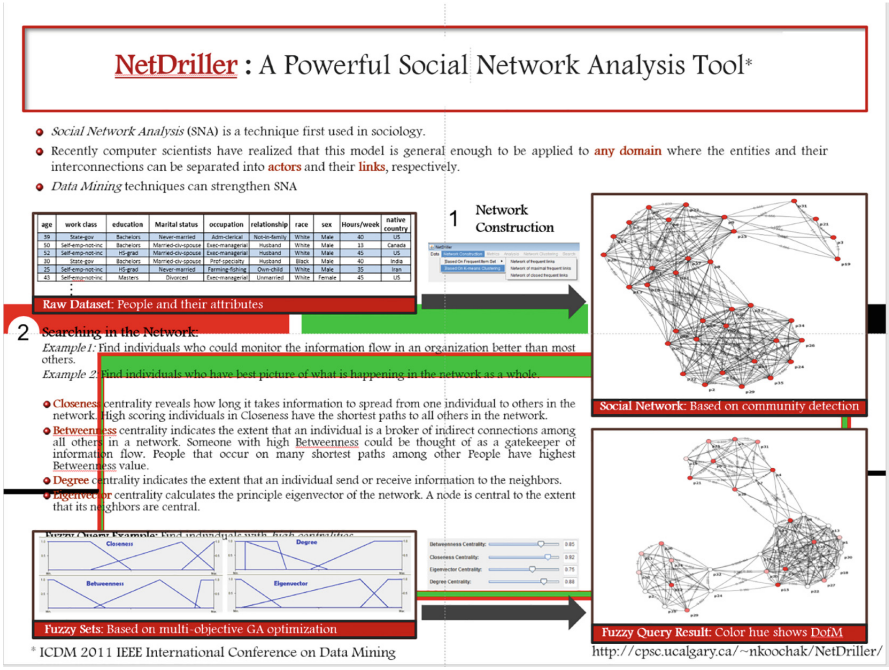

Fig. 2. Basic characteristics of NetDriller.

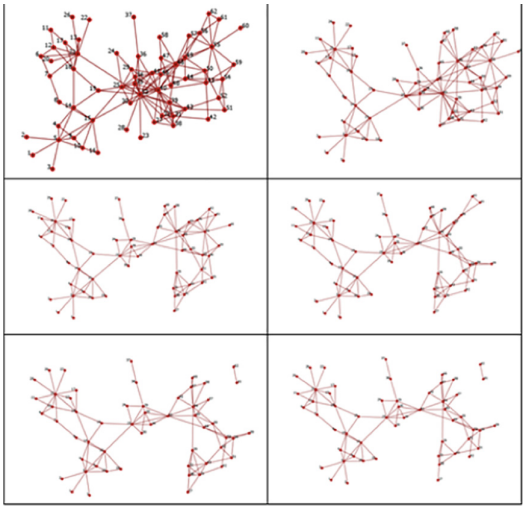

Fig. 3. Sep. 11 network changes after excluding each level nodes for eigenvector centrality measure.

We utilized NetDriller to analyze September 11 terror network (Fig. 3). It is surprising to realize that those who planned for the attacks considered all difficulties they could face. In other words, the network continues to be connected after removing terrorists who were identified as leaders down to level 6. The same is not true when the network of Madrid attacks was analyzed. The latter network became disperse only after removing second level leaders as shown in Fig. 4.

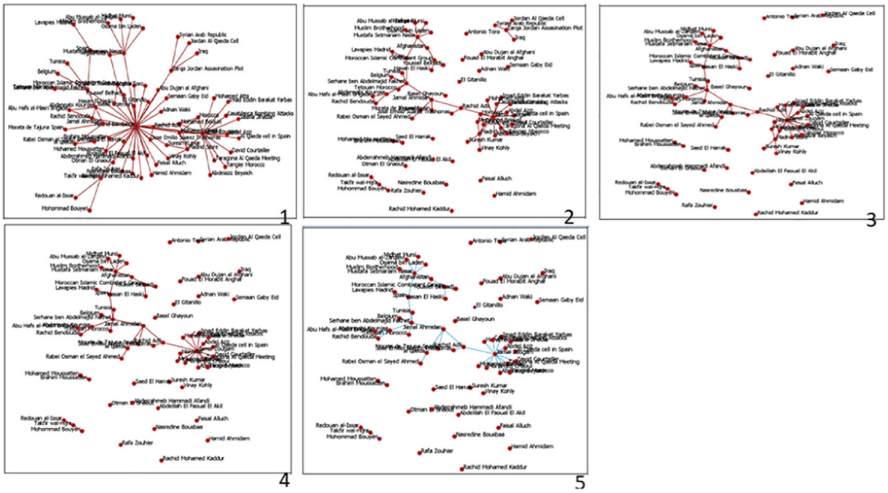

Fig. 4. Madrid network changes after excluding each level nodes for eigenvector centrality measure

Genes	Keywords
PCAP	regulate
HPC5	interact
MAD1L1	activate
HPC4	suppress
HIP1	prognostic
MSR1	biomarker
KLF6	network
PTEN	prognosis
MXI1	elucidate mechanism
CD82	prostate pathway
BRCA2	
CDH1	
ZFHX3	
ELAC2	
HPCQTL19	
HPC3	
CHEK2	
HPC6	
AR	

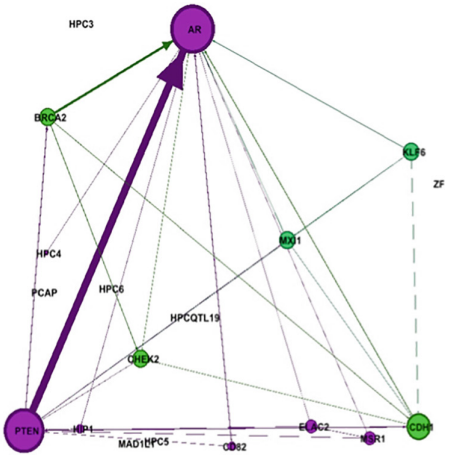

Fig. 5. A list of genes and part of the gene-gene network related to prostate cancer.

NetDriller was also employed using gene expression data related to prostate cancer to identify proteins attributed to the disease. The main result reported by NetDriller is shown in Fig. 5.

Finally, next are some of our ongoing and planned research activities based on the promising results reported in our already published papers. First, in bioinformatics we are tracking disease evolution: spatial and temporal aspects, drug repositioning, etc. Second, in health Informatics we are working on patient monitoring, referral optimization and prediction, etc. Third, we demonstrated the applicability and effectiveness of sequence analysis and prediction for various domains, including financial (e.g., stock, forex), weather, traffic, energy, etc. Finally, other domains and applications considered by our research recommendation, sentiment analysis, opinion mining, spam detection, homeland security, close monitoring and analysis for early warning, etc.

To sum up, our research efforts described in our papers published in the literature and listed in the bibliography illustrate how data mining and network analysis are powerful techniques for data analysis. Further, it is possible to analyze huge amounts of data using limited computing resources, and to develop integrated solutions by combining various aspects leading to robust framework. We have succeeded in developing some techniques from scratch and we also expanded some existing techniques to produce working solutions for our industrial and academic partners. We could help in sophisticated data analysis to maximize knowledge discovery for informative decision making.

References

1. Manber, U., Myers, G.: Suffix arrays: a new method for on-line string searches. In Proceedings of the First Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 319–327 (1990)
2. Cormode, G., Hadjieleftheriou, M.: Methods for finding frequent items in data streams. VLDB J. (2009). https://doi.org/10.10007/s00778-009-0172-z. Unpaginated
3. Boyer, S., Moore, J.: A fast majority vote algorithm. Technical report ICSCA-CMP-32, Institute for Computer Science, University of Texas (1981)
4. Demaine, E.D., López-Ortiz, A., Munro, J.I.: Frequency estimation of internet packet streams with limited space. In: Möhring, R., Raman, R. (eds.) ESA 2002. LNCS, vol. 2461, pp. 348–360. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45749-6_33
5. Karp, R., Papadimitriou, C., Shenker, S.: A simple algorithm for finding frequent elements in sets and bags. ACM Trans. Database Syst. **28**, 51–55 (2003)
6. Manku, G., Motwani, R.: Approximate frequency counts over data streams. In: International Conference on Very Large Data Bases, pp. 346–357 (2002)
7. Metwally, A., Agrawal, D., Abbadi, A.E.: Efficient computation of frequent and top-k elements in data streams. In: International Conference on Database Theory (2005)
8. Greenwald, M., Khanna, S.: Space-efficient online computation of quantile summaries. In: ACM SIGMOD International Conference on Management of Data (2001)
9. Bandi, N., Metwally, A., Agrawal, D., Abbadi, A.E.: Fast data stream algorithms using associative memories. In: ACMSIGMOD International Conference on Management of Data (2007)
10. Alon, N., Matias, Y., Szegedy, M.: The space complexity of approximating the frequency moments. In: ACM Symposium on Theory of Computing, pp. 20–29 (1996). Journal version in J. Comput. Syst. Sci. **58**, 137–147 (1999)
11. Cormode, G., Muthukrishnan, S.: An improved data stream summary: the count-min sketch and its applications. J. Algorithms **55**(1), 58–75 (2005)
12. Xylogiannopoulos, K.F., Karampelas, P., Alhajj, R.: Real time early warning ddos attack detection. Int. J. Cyber Warf. Terror. **7**(3), 44–54 (2017)
13. Üçer, S., Koçak, Y., Ozyer, T., Alhajj, R.: Social network analysis-based classifier (SNAc): a case study on time course gene expression data. Comput. Methods Program. Biomed. **150** (3), 73–84 (2017)
14. Aksac, A., Ozyer, T., Alhajj, R.: Complex networks driven salient region detection based on superpixel segmentation. Pattern Recogn. **66**, 268–279 (2017)
15. Jurca, G., Addam, O., Aksac, A., Gao, S., Ozyer, T., Demetrick, D., Alhajj, R.: Integrating text mining, data mining, and network analysis for identifying genetic breast cancer trends. BMC Res. Notes **9**(1), 236 (2016)
16. Xylogiannopoulos, K.F., Karampelas, P., Alhajj, R.: Repeated patterns detection on big data using classification and parallelism on LERP reduced suffix arrays. Appl. Intell. **45**(3), 567–597 (2016)
17. Ozsoy, M.G., Polat, F., Alhajj, R.: Making recommendations by integrating information from multiple social networks. Appl. Intell. (2016). https://doi.org/10.1007/s10489-016-0803-1
18. Addam, O., Chan, A., Hoang, W., Alhajj, R., Rokne, J.: Foreign exchange data crawling and analysis for knowledge discovery leading to informative decision making. Knowl. Based Syst. **102**, 1–19 (2016)

19. Chen, A., Elhajj, A., Gao, S., Afra, S., Sarhan, A., Kassem, A., Alhajj, R.: Approximating the maximum common subgraph isomorphism problem with a weighted graph. Knowl. Based Syst. **85**, 265–276 (2015)
20. Shafiq, O., Alhajj, R., Rokne, J.G.: On personalizing web search using social network analysis. Inf. Sci. **314**, 55–76 (2015)
21. Xylogiannopoulos, K.F., Karampelas, P., Alhajj, R.: Analyzing very large time series using suffix arrays. Appl. Intell. **41**(3), 941–955 (2014)
22. Rahmani, A., Chen, A., Sarhan, A., Jida, J., Rifaie, M., Alhajj, R.: Social media analysis and summarization for opinion mining: a business case study. Soc. Netw. Anal. Min. **4**, 171 (2014)
23. Xylogiannopoulos, K.F., Karampelas, P., Alhajj, R.: Experimental analysis on the normality of pi, e, phi and square root of 2 using advanced data mining techniques. Exp. Math. **23**(2), 105–128 (2014)
24. Rahmani, A., Afra, S., Zarour, O., Addam, O., Aljomai, R., Koochakzadeh, N., Kianmehr, K., Alhajj, R.: Graph-based approach for outlier detection in sequential data and its application on stock market and weather data. Knowl. Based Syst. **61**, 89–97 (2014)
25. Almansoori, W., Addam, O., Zarour, O., Sarhan, A., Elzohbi, M., Kaya, M., Rokne, J., Alhajj, R.: The power of social network construction and analysis for knowledge discovery in the medical referral process. J. Organ. Comput. Electron. Commer. **24**(2–3), 186–214 (2014)
26. Qabaja, A., Alshalalfa, M., Alanazi, E., Alhajj, R.: Prediction of novel drug indications using a network driven biological data prioritization and integration. J. Cheminform. **6**(1), 1 (2014)
27. Peng, P., Addam, O., Elzohbi, M., Özyer, S., Elhajj, A., Gao, S., Liu, Y., Özyer, T., Kaya, M., Ridley, M., Rokne, J., Alhajj, R.: Analyzing alternative clustering solutions by employing multi-objective genetic algorithm and conducting experiments on cancer data. Knowl. Based Syst. **56**, 108–122 (2014)
28. Kaya, M., Alhajj, R.: Development of multidimensional academic information networks with a novel data cube based modeling method. Inf. Sci. **265**, 211–224 (2014)
29. Rasheed, F., Alhajj, R.: A framework for periodic outlier pattern detection in time series. IEEE Trans. Syst. Man Cybern. **44**(5), 569–582 (2014)
30. Polash Paul, P., Gavrilova, M., Alhajj, R.: Decision fusion for multimodal biometrics using social network analysis. IEEE Trans. Syst. Man Cybern. **44**(11), 1522–1533 (2014)
31. Szeto, J., Lycett, A., Yi, X., Afra, S., Sarhan, A., Xylogiannopoulos, K.F., Karampelas, P., Alhajj, R.: Integrating data mining techniques into a user-friendly framework for visualization of health indicators. Health Inform. **3**, 63 (2014)
32. Alshalalfa, M., Alhajj, R.: Integrating protein networks for identifying cooperative miRNA activity in disease gene signatures. BMC Bioinform. **14**(Suppl. 12), S1 (2013). https://doi.org/10.1186/1471-2015-14-S12-S1
33. Öztürk, O., Aksaç, A., Elsheikh, A.M., Özyer, T., Alhajj, R.: A consistency-based feature selection method allied with linear SVMs for HIV-1 protease cleavage site prediction. PLoS ONE **8**(8), e63145 (2013)
34. Guerbas, A., Addam, O., Zarour, O., Nagi, M., Elhajj, A., Ridley, M., Alhajj, R.: Effective web log mining and online navigational pattern prediction. Knowl. Based Syst. **49**, 50–62 (2013)
35. Almansoori, W., Gao, S., Jarada, T.N., Elsheikh, A.M., Murshed, A.N., Jida, J., Alhajj, R., Rokne, J.: Link prediction and classification in social networks and its application in healthcare and systems biology. Netw. Model. Anal. Health Inform. Bioinform. **1**(1–2), 27–36 (2012)

36. Nagi, M., Elhajj, A., Addam, O., Qabaja, A., Zarour, O., Jarada, T., Gao, S., Jida, J., Murshed, A., Suleiman, I., Özyer, T., Ridley, M., Alhajj, R.: Robust framework for recommending restructuring of websites by analyzing web usage and web structure data. J. Bus. Intell. Data Mining **7**(1/2), 4–20 (2012)
37. Adnan, M., Nagi, M., Kianmehr, K., Ridley, M., Alhajj, R., Rokne, J.: Promoting where, when and what?: an analysis of web logs by integrating data mining and social network techniques to guide eCommerce business promotions. Soc. Netw. Anal. Min. **1**, 173–185 (2012)
38. Rasheed, F., Adnan, M., Alhajj, R.: Out-of-core detection of periodicity from sequence databases. Knowl. Inf. Syst. **36**(1), 277–301 (2013)
39. Khabbaz, M., Kianmehr, K., Alhajj, R.: Employing structural and textual feature extraction for semi-structured document classification. IEEE Trans. Syst. Man Cybern. C **42**(6), 1566–1578 (2012)
40. Rasheed, F., Alhajj, R.: Periodic pattern analysis of non-uniformly sampled stock market data. Intell. Data Anal. **16**(6), 993–1011 (2012)
41. Gao, S., Zeng, J., ElSheikh, A.M., Naji, G., Alhajj, R., Rokne, J., Demetrick, D.: A Closer look at "social" boundary genes reveals knowledge to gene expression profiles. Curr. Protein Pept. Sci. **12**(7), 602–613 (2011)
42. Adnan, M., Alhajj, R.: A bounded and adaptive memory-based approach to mine frequent patterns from very large databases. IEEE Trans. Syst. Man Cybern. B **41**(1), 154–172 (2011)
43. Rasheed, F., Alshalalfa, M., Alhajj, R.: Efficient periodicity mining in time series databases using suffix trees. IEEE Trans. Knowl. Data Eng. **23**(1), 79–94 (2011)
44. Alshalalfa, M., Özyer, T., Alhajj, R., Rokne, J.: Discovering cancer biomarkers: from DNA to communities of genes. Int. J. Netw. Virtual Organ. **8**(1/2), 158–172 (2011)
45. Rasheed, F., Alhajj, R.: STNR: a suffix tree based noise resilient algorithm for periodicity detection in time series databases. Appl. Intell. **32**(3), 267–275 (2010)
46. Rasheed, F., Alshalalfa, M., Alhajj, R.: Adaptive machine learning technique for periodicity detection in biological sequence. J. Neural Syst. **19**(1), 11–24 (2009)
47. Adnan, M., Alhajj, R.: DRFP-tree: disk-resident frequent pattern tree. Appl. Intell. **30**(2), 84–97 (2009)
48. Kaya, M., Alhajj, R.: Multi-objective genetic algorithms based automated clustering for fuzzy association rules mining. J. Intell. Inf. Syst. **31**(3), 243–264 (2008)
49. Kaya, M., Alhajj, R.: Online mining of fuzzy multidimensional weighted association rules. Appl. Intell. **29**(1), 13–34 (2008)