International Journal of Computer Networks and Applications (IJCNA)
DOI: 10.22247/ijcna/2017/41308 Volume 4, Issue 1, January – February (2017)

REVIEW ARTICLE

Big Data Analysis for M2M Networks: Research Challenges and Open Research Issues

Gurkan Tuna
Department of Computer Programming, Trakya University, Edirne, 22020, Turkey.
gurkantuna@trakya.edu.tr

Resul Das
Department of Software Engineering, Fırat University, Elazığ, 23119, Turkey.
rdas@firat.edu.tr

B. Ramakrishnan
Department of Computer Science and Research Centre, S.T. Hindu College, Nagercoil, Tamilnadu, India.
ramsthc@gmail.com

Yilmaz Kilicaslan
Department of Computer Engineering, Adnan Menderes University, Aydın, 09010, Turkey.
yilmaz.kilicaslan@adu.edu.tr

Published online: 14 February 2017

Abstract – In recent years, solutions based on machine-to-machine (M2M) communications have started to support us in many areas of our life and work. However, the amount of data collected by M2M has increased tremendously and surpassed our expectations. This makes it necessary to investigate data mining methodologies and machine learning techniques in order to efficiently utilize large amounts of data gathered by M2M devices. In this paper, we first review existing data mining and machine-learning techniques specifically designed and proposed for M2M networks. Then, we discuss Big Data concept, investigate Big Data analysis techniques, and the importance of Big Data for M2M networks. Finally, we investigate research challenges and open research issues in M2M to provide an insight into future research opportunities.

Index Terms – Machine-to-Machine (M2M), Machine Learning, Data Mining, Big Data.

1. INTRODUCTION

Machine-to-Machine (M2M) is a broad label that refers to direct communication between devices using a wireless or wired communication channel [1-3]. Thanks to M2M technologies, devices can exchange information with each other across remote sites in an automatic way without human intervention. In M2M communications, one or more remote sensor nodes collect data and then send it to a network, periodically or immediately depending on the policy, where it is next routed to a central server. Afterwards, the software application running on the central server analyses the collected data and carries out one or more predetermined tasks. Different from telemetry systems used in the past, sensor technology used in M2M applications offers increased accuracy and sensitivity. In addition, Today's computers used in data analysis have higher processing power and memory than their predecessors, and thus software applications run faster.

At the beginning, M2M was realized through one-to-one connections on fixed networks. However, nowadays M2M services are mostly deployed over wireless networks due to the prevalence of mobile networks. Although many M2M deployments make use of proprietary or short-range radio links, some M2M deployments which require mobility or require high data transfer rates make use of cellular frequencies [2-3]. Cellular-based M2M deployments have advantages for short-term deployments in terms of installation and provision [4]. Since there are many standards to address communication needs and there is a degree of fragmentation, standardization efforts are ongoing to make it possible to deliver practical and cost-effective M2M solutions [5].

M2M solutions enable to collect all kinds of information from a few bytes to several megabytes depending on the application requirements. Different from the traditional M2M applications, at the moment, timely data can be provided by M2M services to the users. M2M has not only led to the start of innovative solutions and new business opportunities in different sectors but also reduced man power needs and operational costs. Furthermore, M2M improves the quality of provided services. Therefore, recently, M2M has gained a great momentum as one of the key enabling technologies for a broad range of applications from the traditional ones including remote

REVIEW ARTICLE

monitoring, industrial automation, inventory tracking, automated billing, and security, traffic control to the emerging ones such as e-health, smart grid, smart home and smart cities [2, 3, 6]. M2M's bright future lies in the fact that it is a flexible technology using common equipment in new ways [60].

In the past data collection by M2M applications was mostly limited to individual activities, whereas now it is one of the reasons behind the global increase in data quantities. Therefore, although conventional data mining methodologies and machine learning techniques have played important roles in achieving the potential gains of many industrial and non-industrial applications, they are not sufficient anymore for handling the tremendous amount of data provided by M2M applications [7, 8]. As shown in Figure 1, handling of obtained data is one of the key stages of current M2M applications. Since the real potential of M2M services lies in looking at data streams in real time and identifying relevant patterns as early as possible, one of the key features of current M2M applications is seen to be their ability to deal with streaming data. In this way, problems are expected to be detected before they arise in the ongoing processes.

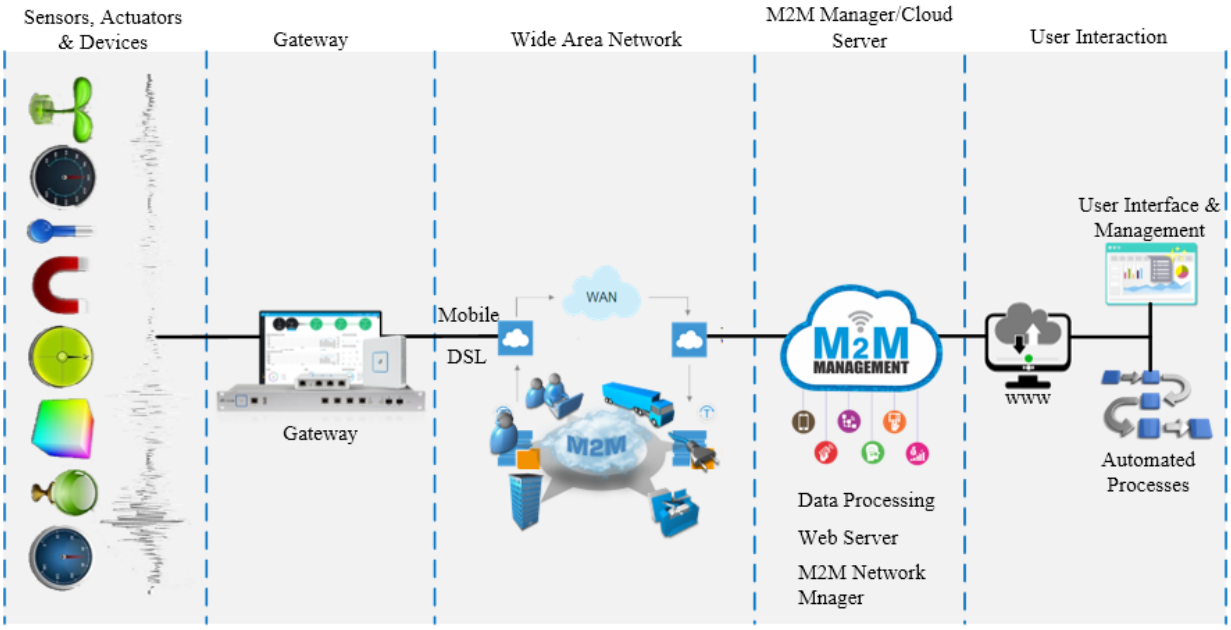

Figure 1 Overall illustration of a typical M2M application.

When data is generated with a rate and volume at a scale beyond the capacity of the state-of-the-art systems and therefore creates a need for revolutionary solutions, it is called Big Data [9-11, 56-59]. With the revolutionary Big Data solutions, the tremendous amount of data generated by M2M applications can be handled effectively. This paper offers a survey of Big Data analysis solutions for M2M networks. The rest of the paper is organized as follows. Section 2 is devoted to the presentation of data mining methodologies and machine learning techniques. Section 3 presents the impact of Big Data on M2M solutions. Research challenges are given in Section 4. Open research issues are discussed in Section 5. The paper is concluded in Section 6.

2. DATA MINING AND MACHINE LEARNING FOR M2M

M2M promises to revolutionize the way we work and live by delivering social and environmental benefits in many sectors. It improves the economic efficiency of all organizations. It allows for a shift to autonomous operation from manual operation and thereby offers better environmental sustainability. Finally, it provides better organizational productivity and efficiency by the rate of productivity and the speed of decision making processes. However, considering the increase in data amounts, it is clear that all these benefits can be considered effective only if the right data is processed quickly and at the right time. Therefore, the role that data mining methodologies and machine learning techniques play in data processing and management is in this respect.

The main purpose of data mining is to extract relationships, which have not previously been discovered, and summarize them into useful information for various goals. [12, 13]. Machine learning algorithms generate out of these relationships computer programs that can detect them in new data. Both data mining and machine learning involve pattern recognition capabilities relating to the discovery and characterization of patterns in high dimensional data. Measurable features extracted from the data are used to identify

REVIEW ARTICLE

patterns. Machine Learning mas mainly to do with the design and development of algorithms, which enable computers/autonomous systems to learn new behaviours resting on a set of empirical data. This will allow for the design of intelligent systems, which can automatically recognize complex patterns and thereby make decisions based on data [14-16]. Both data mining and machine learning provide techniques to deal with the emerging Big Data problem.

Being a special technique in knowledge discovery, data mining can be viewed as an iterative and interactive process concerned with uncovering statistically significant structures and events, patterns, associations, and anomalies in data [17-19]. In the overall data mining processes can be divided into three main stages, namely data pre-processing, pattern recognition, and result interpretation, as shown in Figure 2 [20]. Each of these stages covers a few steps.

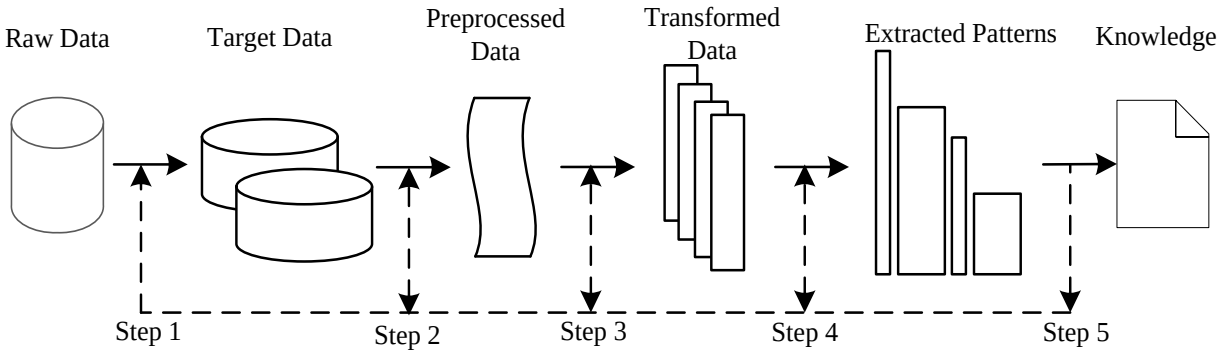

Figure 2 Data mining process.

The data preprocessing stage covers steps 1-3. The pattern recognition stage covers step 4. The result interpretation stage covers step 5. The operations included in these steps are:

- Step 1: Data fusion, sampling, and multi-resolution analysis;
- Step 2: Noise reduction, feature extraction, and normalisation;
- Step 3: Dimension reduction;
- Step 4: Classification and clustering;
- Step 5: Visualisation and validation.

Machine Learning is a field of computer science, probability theory and optimization theory, and, as already stated, relates to the design, development and implementations of the algorithms which give computers the capability to learn and find hidden patterns without being explicitly programmed [21, 22]. In addition, it provides solutions for sophisticated problems for which a logical and/or procedural approach would be impossible or unfeasible [21, 22]. Besides, it automates analytical model building [21, 23]. Relying on iterative approaches, machine-learning models can independently adapt to and learn from previous computations to produce reliable decisions and results. There are many machine-learning algorithms, such as Bayesian estimation methods, Artificial Neural Networks (ANNs), Random Forests (RFs), and Support Vector Machines (SVMs) [21-24]. Although in some cases some algorithms may work better than others, it is very rare that one algorithm strictly dominates all others in tackling a given machine learning problem.

3. BIG DATA FOR M2M NETWORKS

Big Data is one of the most excited and popular terms used in data analytics to describe a set of methodologies, techniques, and technologies which use new integration forms to uncover hidden patterns from massive scale, diverse, and complex datasets. Addressing the leading challenges of statistical science, it serves to broaden not only the algorithmic but also the theoretical perspective [29, 30, 40]. It always involves massive data, including data streams and data heterogeneity. In addition, since its major key features are multiple sources, huge volume, and fast-changing nature, it is difficult for commonly used traditional computing methods such as machine learning, information retrieval and data mining to efficiently support the processing, analysis and computation of Big Data [40]. Therefore, in recent years, statistical methods such as clustering methods, linear regression models and bootstrapping schemes have been adapted to process Big Data [29]. Generally, Big Data can be classified based on the five main aspects of data and its content: source, store, format, staging, and processing [11, 39]. Each specific aspect of Big Data can be categorized as follows:

- Source:
  - Machine;
  - Sensing;
  - The Internet of Things;

REVIEW ARTICLE

  - Web;
  - Transactions.
- Format:
  - Unstructured;
  - Semi-structured;
  - Structured.
- Processing:
  - Batch;
  - Real-time.
- Staging:
  - Cleaning;
  - Normalisation;
  - Transformation.
- Store:
  - Key/Value-based;
  - Column-based;
  - Document-based;
  - Graph-based.

M2M communications rely on a large number of interconnected physical objects, which create mesh of machines and produce massive volume of data [54]. Hence, M2M networks may sometimes carry tremendous data gathered from a large number of M2M devices. Since Big Data, technology allows analysing very large, fast and heterogeneous data streams which cannot be effectively and/or affordably handled using traditional data management approaches and tools, it can play a key role in the analysis of M2M data [55].

4. RESEARCH CHALLENGES

Data mining requires intensive statistical computations on large datasets in widely-deployed big M2M applications. In addition, the heterogeneous and geographically dispersed nature of some M2M deployments necessitates distributed data mining, and poses challenges in terms of data integrity, privacy and security. Furthermore, in some M2M applications, it is necessary to scale up for high speed data streams due to the increase in data transfer rates. Similar to data mining operations, machine-learning applications in M2M exhibit various challenges. When applying machine-learning techniques for M2M communication unexpected problems may arise due to the complexity and processing time added by the scale of applications. The large scale makes trivial operations costly in terms of computation time and memory requirement and forces the system designers to reconsider the applied machine learning algorithms since the cost of learning arises primarily from bandwidth and disk reads [25, 40]. In this section, a literature review is presented to discuss the challenges in data mining and machine learning for M2M applications.

Certain concerns must be addressed when applying large-scale data mining in M2M domains. Format-related problems create serious challenges in the extraction of features [26, 54]. Random sampling techniques and traditional clustering approaches may not be applied effectively in knowledge discovery processes of M2M applications. In addition, in most cases, the data mining process needs to be iterated several times in order to obtain a reasonably sized training set. This need arises due to the scarcity of labelled data in the classification problem. Furthermore, some M2M applications need data fusion approaches to mine the data collected by different sensors with different resolutions.

Large-scale data mining is indeed a source of many open research problems [52, 53]. Before applying pattern recognition algorithms, key features from large, complex and sometimes multi-dimensional data must be extracted. However, the features must be relevant to the problem and not be sensitive to minor changes and variant to translation, rotation, and scaling [27]. Nevertheless, existing algorithms for data mining are generally not scalable and robust enough to be used in large-scale M2M applications. In addition, due to the massive data volume, the implementation of existing data mining algorithms on large-scale distributed systems can speed up exploration and analysis of data.

In large-scale M2M applications, large volumes of data are generated as streams, which must be analyzed online as soon as they arrive. In this regard, handling streaming data is one of the key tasks for Big Data analysis. However, mining big data streams in M2M applications faces three main challenges resulting from the factors of volume, volatility, and velocity. Speaking of volume and velocity, a huge volume of data must be processed in a short time. However, during this process, the amount of available data continually increases. Therefore, it is necessary to use incremental approaches, which allows for immediately incorporating newly coming information and to rely on online processing techniques if all data cannot be kept in memory [28]. In some M2M applications, due to the volatility resulting from the dynamic environment with changing patterns, old data cannot be re-processed at a later time. To conclude, in M2M applications, data mining of big data streams is affected by multiple factors in multiple ways.

In almost all machine-learning tasks, feature engineering and feature selection play an important part [49]. Even if sophisticated estimation algorithms are developed and powerful computing capabilities are available, significant time and energy are spent on looking over the data itself with the goal of identifying additional information, which may be in the

REVIEW ARTICLE

features already included. That is, feature engineering aims at taking the maximum amount of useful information out of a given set of input data.

Most problems in an M2M domain might be approached at least from two different angles: from the point of view of learning from unlabeled data and from the point of view of active learning. In a paradigm of learning from unlabeled data, machine-learning algorithms must do with a limited amount of labelled data and capitalize on unlabeled data with semi-supervised learning methods. As for active learning, this requires machine-learning algorithms to typically place a limited number of queries to get labels [50]. Here, the goal is to optimize the queries to label data.

Researches on machine learning are often carried out in vitro using publicly available datasets and they are far from motivating practical applications [27]. Although such approaches are fruitful for academic activities, it is not effective for real-world scenarios such as M2M and cannot effectively address unforeseen problems specific to practical M2M applications. Moreover, M2M platforms handling Big Data might not function properly in accordance with the goals of their operators. Therefore, specific requirements of M2M applications may not be addressed efficiently in terms of data quality, database utilisation, processing load, computational efficiency, feature extraction and analysis applications [54]. Therefore, it is crucial to work with processing and memory efficient system architectures, which meet such requirements.

5. OPEN RESEARCH ISSUES

As data, mining is an applied discipline and widespread used in many application domains, it requires close cooperation with the designers. Several research issues remain open in this area, concerning the theory, the framework adopted, the domains of application, the system and programming languages, and the algorithmic approaches to be taken. Likewise, machine learning is one of the hot topics in the last decades and comes with widely used techniques in real-world M2M applications. Although the amount of work performed by machine learning applications in large-scale M2M applications has increased significantly, there is a number of open research issues in many domains, especially in terms of speed, efficiency, parallelism and data streaming. In this section, we present open research issues on the use of data mining and machine learning in M2M applications.

While the mainstream M2M data mining research focuses on the innovation and deployment of algorithms and tools, the workable capability of the algorithms and the tools in real world M2M scenarios has been mostly neglected. Many research issues in data mining field need to be studied to effectively address the main requirements of M2M applications. Similar to traditional data mining applications, mining in-depth data patterns and/or structured knowledge in unstructured data are the typical open research issues in M2M domain.

Although predictive modelling for data streams has received considerable attention in the last years, many of the approaches have overlooked major challenges imposed by real-world M2M applications. Having control in the full cycle of knowledge discovery, exploring data interactively through parallel implementation [41], categorizing multiple data sources and types, reducing dimension to handle multi-dimensional data, dealing with legacy systems found in M2M services, protecting data confidentiality, dealing with delayed and/or incomplete information, development and evaluation of data stream mining and approximation algorithms, data staging, scalable algorithms for classification and clustering, and complex data analysis remain major open research issues in this area. Last but not least, applied statistics should be widely used in data mining applications for M2M communications to ensure that the conclusions drawn from the large-scale data are statistically significant.

As for machine learning in M2M applications, Support Vector Machines (SVMs) appear to be very effective learning models to perform a non-linear classification by implicitly mapping their inputs into high-dimensional feature spaces using the so-called kernel trick. This allows for the memory consumption and training runtime of an SVM application to scale with the size of the data set [42]. Another method to be used for improving M2M communications is Stochastic Gradient Descent (SGD), which approximates the gradient descent optimization that makes use of a random process to allow machine-learning algorithms to converge faster [43]. As training SVMs on large M2M datasets is hard, combining SVMs and SGD may achieve satisfactory results on large datasets [44]. In addition, machine-learning algorithms can be distributed across multiple computers or cores and run, and in this way, the operation speed can be increased [38]. These are open research issues that have not been investigated thoroughly.

Deep machine learning, an emerging field that promises unparalleled results on various data analysis problems, also needs to be discussed and analyzed to a further extent [34, 35]. Deep machine learning uses pre-training on unlabeled data to learn the best features for the next layer and leads our focuses to another research issue called convex Non-negative Matrix Factorization (NMF), a kind of auto-encoder [36, 37]. In this respect, making best predictions and selecting the smallest possible subset of relevant input variables plays a key role for feature selection. In addition to this, better training algorithms, regularization and activation functions for deep learning can also be useful. While research efforts in this field attempt to create better representations and develop models to acquire these representations from large-scale unlabeled data and some have produced state-of-the-art results on various tasks, the

REVIEW ARTICLE

application of deep learning architectures in M2M applications is an open research issue. In addition, although reliable semi-supervised learning in which a small set of labelled data is complemented by a large set of unlabeled data has often produced poor performance in practice [48], it is known that unlabeled data can aid the learning process and the efficiency of semi-supervised learning in large-scale M2M applications is still a mystery.

Recently Big Data has been one of the most heavily investigated topics in many application domains including M2M. Big Data has many consequences from algorithmic and theoretical viewpoints. Generally, in addition to involving massive data, it includes data heterogeneity and data streams. Some statistical techniques such as clustering methods, linear regression models, and bootstrapping schemes have been successfully implemented in Big Data solutions. Random forests which are a powerful nonparametric statistical approach that allow considering regression problems as well as multiclass classification problems in a single framework generally exhibit superior performance with only a few selected parameters to tune [29, 31-33]. Therefore, adapting random forests to M2M Big Data solutions is another open research issue.

In machine learning, probabilistic programming is a programming paradigm to manage uncertain information [45]. Its main goal is to facilitate the construction of machine learning applications by using probabilistic programming. As shown in the literature, in addition to making it easy to build effective machine learning applications, it has many other benefits. The power of probabilistic programming comes from the inference algorithm used. Basically, on the basis of new sets of training data, the inference algorithm continuously readjusts probabilities [51].

Calculating necessary computational resources for machine learning problems is a difficult task, because over-budgeting on powerful computing systems can waste significant money, but under-budgeting can produce severe bottlenecks in model construction and deployment [46]. In this respect, cloud computing can be quite useful for making computational pipelines more expandable. Cloud computing allows for the deployment of larger virtual machines or greater numbers of machines working in parallel with relatively low cost and high speed and brings many advantages to M2M [47]. In this way, it becomes much easier to budget appropriately when setting up a machine learning system, especially when working with very large M2M data sets.

6. CONCLUSION

Thanks to M2M, the technology enabling the exchange of data between any kinds of devices in an autonomous way and the widespread use of wireless communications technologies, remote data acquisition and control is nowadays available in a more cost-effective way and this offers many benefits to the industry. Expectedly, data quantities gathered by M2M devices are growing constantly as M2M becomes increasingly widespread, and M2M is one of the leading drivers of the Big Data trend. As it is well known, raw data gathered by M2M devices does not serve the purpose of drawing conclusions about the information. Hence, data mining and machine learning techniques are used in many industries to allow companies and organisations to make effective business decisions. In this respect, as data assets from M2M solutions play a key role, analysing them helps to gain deeper insights into business workflows and production processes. In this paper, existing data mining methodologies and machine learning techniques designed for M2M have been reviewed. In addition, data acquisition and processing-related research challenges and open research issues in M2M have been investigated, and Big Data concept and its role in M2M applications has been discussed.

REFERENCES

[1] J. N. Al-Karaki, K. -C. Chen, G. Morabito and J. de Olieveira, "From M2M communications to the Internet of Things: Opportunities and challenges," *Ad Hoc Networks*, vol. 18, pp. 1-2, 2014.
[2] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. D. Johnson, "M2M: From Mobile to Embedded Internet," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 36-43, April 2011.
[3] J. Holler, V. Tsiatsis, C. Mulligan, S. Avesand, S. Karnouskos and D. Boyle, "From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence," Academic Press: MA, USA, 2014.
[4] A. Biral, M. Centenaro, A. Zanella, L. Vangelista, and M. Zorzi, "The challenges of M2M massive access in wireless cellular networks," *Digital Communications and Networks*, vol. 1, no. 1, pp. 1-19, February 2015.
[5] ETSI, "Machine to Machine Communications," TS 102 689.
[6] D. Boswarthick, O. Elloumi, and O. Hersent (Eds), M2M Communications: A Systems Approach, Wiley: West Sussex, UK, 2012.
[7] Z. Fan, Q. Chen, G. Kalogridis, S. Tan, and D. Kaleshi, "The power of data: Data analytics for M2M and smart grid," *2012 3rd IEEE PES international conference and exhibition on Innovative Smart Grid Technologies (ISGT Europe)*, pp. 1-8, 2012.
[8] G. Suciu, A. Vulpe, A. Martian, S. Halunga, and D. N. Vizireanu, "Big Data Processing for Renewable Energy Telemetry Using a Decentralized Cloud M2M System," *Wireless Personal Communications*, 2015.
[9] G. Suciu, V. Suciu, A. Martian, R. Craciunescu, A. Vulpe, I. Marcu, S. Halunga, and O. Fratu, "Big Data, Internet of Things and Cloud Convergence – An Architecture for Secure E-Health Applications," *Journal of Medical Systems*, vol. 39, no. 141, 2015.
[10] A. J. Jara, D. Genoud and Y. Bocchi, "Big Data for Cyber Physical Systems: An Analysis of Challenges, Solutions and Opportunities," *2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, pp. 376-380, 2014.
[11] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of "big data" on cloud computing: Review and open research issues," *Information Systems*, vol. 47, pp. 98-115, 2015.
[12] M.-S. Chen, J. Han, and P. S. Yu, "Data Mining: An Overview from a Database Perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 866-883, December 1996.
[13] D. Talia, P. Trunfio, and F. Marozzo, "Introduction to Data Mining," in *Data Analysis in the Cloud*, pp. 1-25, 2016.
[14] I. Kononenko and M. Kukar, "Machine Learning Basics," in *Machine Learning and Data Mining*, pp. 59-105, 2007.

REVIEW ARTICLE

[15] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881-892, July 2002.
[16] G. Holmes, A. Donkin, and I. H. Witten, "WEKA: a machine learning workbench," *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems*, pp. 357-371, 1994.
[17] Z. He, "An overview of data mining," in *Data Mining for Bioinformatics Applications*, pp. 1-10, 2015.
[18] A. Lausch, Andreas Schmidt, and Lutz Tischendorf, "Data mining and linked open data – New perspectives for data analysis in environmental research," *Ecological Modelling*, vol. 295, pp. 5-17, 2015.
[19] C. C. Aggarwal, *Data Mining*, Springer International Publishing: New Delhi, India, May 2015. ISBN: 978-3-319-14141-1.
[20] Sapphire: Large Scale Data Mining and Pattern Recognition. Available at: http:// computation.llnl.gov/casc/sapphire/overview/overview.html. Accessed: October 7, 2015.
[21] S. M. Weiss and C. A. Kulikowski, Computer Systems that Learn. Morgan Kaufmann: San Mateo, CA, 1991.
[22] R. Duda and P. Hart, Pattern Recognition and Scene Analysis. Wiley, New York, 1973.
[23] R. E. Abdel-Aal, "Comparison of Algorithmic and Machine Learning Approaches for the Automatic Fitting of Gaussian Peaks," Neural Computing & Applications, vol. 11, no. 1, pp. 17-29, 2002.
[24] P. D. Wasserman, Neural Computing: Theory and Practice. Van Nostrand Reinhold, New York, 1989.
[25] G. Krempl, I. Žliobaite, D. Brzeziński, E. Hüllermeier, M. Last, V. Lemaire, T. Noack, A. Shaker, S. Sievi, M. Spiliopoulou, and J. Stefanowski, "Open challenges for data stream mining research," *SIGKDD Explor. Newsl.*, vol. 16, no.1, pp. 1-10, September 2014.
[26] H. Mannila. Methods and Problems in Data Mining. In Proceedings of the 6th International Conference on Database Theory (ICDT '97), F. N. Afrati and P.G. Kolaitis (Eds.). Springer-Verlag, London, UK, 1997. pp. 41-55.
[27] C. E. Brodley, U. Rebbapragada, K. Small, and B. C. Wallace, "Challenges and Opportunities in Applied Machine Learning," *AI Magazine*, vol. 33, no. 1, pp. 11-24, March 2012.
[28] C. Parker "Unexpected challenges in large scale machine learning," in *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications (BigMine '12)*. ACM, New York, NY, USA, 2012. pp. 1-6.
[29] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, N. Villa-Vialaneix, "Random forests and big data," *47`eme Journ´ees de Statistique de la SFdS*, Jun 2015, Lille, France. 2015.
[30] M. I. Jordan, "On statistics, computation and scalability," *Bernoulli*, vol. 19, no. 4, pp. 1378-1390, 2013.
[31] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
[32] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: a survey and results of new tests," *Pattern Recognition*, vol. 44, no. 2, pp. 330-349, 2011.
[33] A. Ziegler and I. R. König, "Mining data with random forests: current options for real-world applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 1, pp. 55-63, 2014.
[34] I. Arel, D. C. Rose, and T. P. Karnowski, "Research frontier: deep machine learning--a new frontier in artificial intelligence research," *IEEE Computational Intelligence Magazine*, vol. 5, no. 4, pp. 13-18, November 2010.
[35] C. L. Philip Chen and Chung-Yang Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Sciences*, vol. 275, pp. 314-347, 10 August 2014.
[36] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798-1828, 2013.
[37] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 145-154, 2011.
[38] J. X. Dong, A. Krzyzak, and C. Y. Suen, "Fast SVM training algorithm with decomposition on very large data sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 603-618, 2005.
[39] P. Pääkkönen and D. Pakkala, "Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems," *Big Data Research*, vol. 2, pp. 166-186, 2015.
[40] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and Challenges of Big Data Research," *Big Data Research*, vol. 2, pp. 166-186, 2015.
[41] Y. Ma, H. Wu, L. Wang, B. Huang, R. Ranjan, A. Zomaya, and W. Jie, "Remote sensing big data computing," *Future Generation Computing Systems*, vol. 51, pp. 45-60, October 2015.
[42] J. Cervantes, F. G. Lamont, A. López-Chau, L. R. Mazahua, J. Sergio Ruíz, "Data selection based on decision tree for SVM classification on large data sets," *Applied Soft Computing*, vol. 37, pp. 787-798, December 2015.
[43] L. Bottou, "Large-Scale Machine Learning with Stochastic Gradient Descent," in *Proceedings of COMPSTAT'2010*, pp. 177-186, 30 September 2010.
[44] K. Sopyła and P. Drozda, "Stochastic Gradient Descent with Barzilai–Borwein update step for SVM," *Information Sciences*, vol. 316, pp. 218-233, 2015.
[45] T. D. Kulkarni, P. Kohli, J. B. Tenenbaum, and V. Mansinghka, "Picture: A probabilistic programming language for scene perception," *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4390-4399, 2015.
[46] K. Kumar and Yung-Hsiang Lu, "Cloud Computing for Mobile Users: Can Offloading Computation Save Energy?" *Computer*, vol. 43, no. 4, pp. 51-56.
[47] A. Botta, W.de Donato, V. Persico, and A. Pescapé, "Integration of Cloud computing and Internet of Things: A survey," *Future Generation Computer Systems*, 2015.
[48] Y. Yang and X. Liu, "A robust semi-supervised learning approach via mixture of label information," *Pattern Recognition Letters*, vol. 68, part 1, pp. 15-21, 15 December 2015.
[49] Y. Zhu, E. Zhong, Z. Lu, and Q. Yang, "Feature engineering for semantic place prediction," *Pervasive and Mobile Computing*, vol. 9, no. 6, pp. 772-783, December 2013.
[50] R. Hu, B. M. Namee, and S. J. Delany, "Active learning for text classification with reusability," *Expert Systems with Applications*, vol. 45, pp. 438-449, 1 March 2016.
[51] H. Rajaona, F. Septier, P. Armand, Y. Delignon, C. Olry, A. Albergel, and J. Moussafir, "An adaptive Bayesian inference algorithm to estimate the parameters of a hazardous atmospheric release," *Atmospheric Environments*, vol. 122, pp. 748-762, December 2015.
[52] N. V. Chawla, Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook*, Springer US, 2010, pp. 875-886.
[53] J. Gama, R. Sebastião, and P. P. Rodrigues, "Issues in evaluation of stream learning algorithms," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, June 2009, pp. 329-338.
[54] A. Ahmad, A. Paul, and M. M. Rathore, "An efficient divide-and-conquer approach for big data analytics in machine-to-machine communication," *Neurocomputing*, vol 174, part A, pp. 439-453, 22 January 2016.
[55] G. Suciu, A. Vulpe, O. Fratu, and V. Suciu, "M2M remote telemetry and cloud IoT big data processing in viticulture," *2015 International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2015, pp. 1117-1121.
[56] R. T. Kouzes, G. A. Anderson, S. T. Elbert, I. Gorton, and D. K. Gracio, "The changing paradigm of data-intensive computing," *Computer*, vol. 42, no. 1, pp. 26-34, 2009.

REVIEW ARTICLE

[57] I. Gorton, "Software architecture challenges for data intensive computing," in *Seventh Working IEEE/IFIP Conference on Software Architecture (WICSA 2008)*, Feb. 2008, pp. 4-6.
[58] R. Das, D. Demirol, G. Tuna, "A Novel Software Tool for Mining Access Patterns Efficiently from Web User Access Logs", 2nd International Conference on Engineering and Natural Sciences (ICENS), 24-28 Mayıs 2016, Sarajevo (Saraybosna), Bosnia and Herzegovina (Bosna ve Hersek) pp.2836-2843.
[59] R. Das, I. Türkoglu, (2009). "Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method", Expert Systems with Applications (ISI), 36(3), 6635-6644 pp., 2009, DOI: doi:10.1016/j.eswa.2008.08.067.
[60] R. Daş, G. Tuna, (2015), "Machine-to-Machine Communications for Smart Homes", International Journal of Computer Networks and Applications (IJCNA), 2(4), 196-202.

Authors

Gurkan Tuna is currently an Associate Professor at the Department of Computer Programming of Trakya University, Turkey. Dr. Tuna has authored several papers in international conference proceedings and refereed journals, and has been actively serving as an Associate Editor for IEEE Access and Australian Journal of Electrical and Electronics Engineering journals. His current research interests include smart cities, smart grid, wireless sensor networks, underwater networks and M2M communications.

Resul Das received his B.Sc. and M.Sc. in Computer Science from Firat University in 1999, 2002 respectively. Dr. Das received his Ph.D. degree from Electrical and Electronics Engineering Department at the same university in 2008. He is currently an Associate Professor at the Department of Software Engineering of Firat University, Turkey. He has authored several papers in international conference proceedings and refereed journals, and has been actively serving as a reviewer for international journals and conferences. His current research interests include complex networks, computer networks, web mining, knowledge discovery, and information and network security.

B. Ramakrishnan is currently working as Associate Professor in the Department of Computer Science and Research Centre in S. T. Hindu College, Nagercoil, India. He received his M.Sc Degree from Madurai Kamaraj University, Madurai and received M.phil (Com. Sci) from Alagappa University, Karaikudi. He earned his Doctorate degree in the field of computer science from Manonmanium Sundaranar University, Thirunelveli. He has a researching experience of 29 years. He has twelve years of research experience, published more than 50 research articles in reputed international journal. His research interest lies in the field of Vehicular Network, Mobile Network and Communication, Cloud Computing, Ad-hoc Network and Network Security.

Yilmaz Kilicaslan is a professor at the Department of Computer Engineering in Adnan Menderes University, Turkey. He has authored many papers in refereed journals and international conference proceedings and has been actively serving as a reviewer for international journals and conferences.