B9DA103
DATA MINING

BIG DATA MINING PROCESS AND APPLICATION

By

Mayank Ketan Gandhi

10394669

Lecturer

TERRI HOARE

PART A

ARTICLE CRITIQUE
“The CRISP-DM Model: The New Blueprint for Data Mining”
Volume 5 Number 4 Fall 2000
- Colin Shearer

INTRODUCTION

The article, "The CRISP-DM Model: The New Blueprint for Data Mining", by Colin Shearer seeks to address the framework for designing, creating, building, testing, and deploying data mining projects. CRISP-DM – the Cross Industry Standard Process for Data Mining – is the leading approach for managing data mining, predictive analytics and data science projects. While this article provides good focus on the business understanding piece, the model no longer seems to be actively maintained and further the framework itself has not been updated on issues on working with new technologies, such as Big Data. The abstract of the article clearly shows how the CRISP-DM methodology was conceived by Daimler Benz, Integral Solutions Ltd (ISL), NCR, and OHRA. The article breaks the process of the data mining process in six phases. Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. These six phases are very important and helps the company in understanding the data mining process and provides a path to follow while planning and carrying out a data mining project.

Concerns in Using CRISP-DM Model

As we know that the model is no longer being maintained and the industry wants to keep up-to-date with the newest frameworks, and the newest technology. The Industry uses CRISP-DM because it recognizes the need for a repeatable approach. However, there are some persistent problems with how model is generally applied. Technologies like Big Data have various effects on CRISP-DM methodology.

For example, Big Data technologies mean that there can be additional effort spend in the Data Understanding phase as the business tackles with the complexities that are involved in the Big Data sources. Few of the issues on using Crisp-DM model are discussed below

- **Lack of Coherence**

The project team needs to have a good clarity on the business problem so that a solution is proposed keeping all the business objectives in check. Clearly, now they understand the business objective, but they want to reduce overhead and proceed into analysing the data. Often this results in models that don't meet business objectives.

- **Need for Rework**

There is difficulty if the analyst does not have real clarity on the business problem. If the developed model does not seem to meet the business objectives, then there are very less options left. There is a lot of rework on understanding the data or developing a new model rather than re-evaluating and understanding the business objective.

- **Handover to Deployment Phase**

In the Evaluation phase, the analytic teams don't think about deployment and operationalization of their models at all. Whether the model is easy to implement or hard and whether it's usable once deployed is not the analysts' concern. This results in increasing the time and cost of deploying a model and contributes to the huge percentage of models that will not have much business impact.

- **Negligence on Iterations**

While the article describes in detail about the deployment phase, there is no mention or focus on the Iteration process. Professionals know that models need to be kept up to date if they are to continue to be valuable. Everybody knows that the business circumstances can change and undermine the value of a model. If these models are unmonitored and unmaintained and there is no iterative work applied on these models, then this article undermines the long-term value of analytics.

Different Approaches

There are other alternatives that can be considered for data mining methodology. They are discussed below

ASUM-DM
In late 2015, an extended version of CRISP-DM is proposed by IBM Corporation called ASUM-DM (the Analytics Solutions Unified Method). It has the same process as of Crisp-DM but only with one with one additional step. ASUM-DM adds a new deployment/operation wing to CRISP-DM. The development phase stays the same as CRISP-DM however in deployment new facets are added such as collaboration, version control, security, and compliance. CRISP-DM repeats itself in ASUM-DM as the development part however it misses an important step being data validation. It ensures more success for the project and prevents the failures and faults of the data-mining project.

Microsoft Team Data Science Process (TDSP)

The TDSP process model provides a dynamic framework to machine learning solutions that have been through a robust process of planning, producing, constructing, testing, and deploying models. TDSP process has only four phases, Business Understanding, Data Acquisition and understanding, Modelling and Deployment. In TDSP, business objectives are given major importance while proposing the solution. Modelling phase in TDSP ensures that the model adds business value and involves examining key metrics in models. There is a recurrent evaluation of the deployed models, thus iterations to deployed models are done as and when needed. The TDSP is a team-oriented solution that prioritizes teamwork and collaboration throughout the scope of the project.

SEMMA

SEMMA stands for Sample, Explore, Modify, Model, and Assess. It has a list of sequential steps developed by SAS Institute and guides the implementation of data mining application. Although SEMMA is often considered to be a general data mining methodology, SAS claims that it is "rather a logical organization of the functional tool set of" one of their products, SAS Enterprise Miner, "for carrying out the core tasks of data mining". SEMMA mainly focuses on the modelling tasks of data mining projects, leaving the business aspects out unlike CRISP-DM which has a Business Understanding phase.

Knowledge Discovery in Databases (KDD)

KDD refers to the broad process of finding knowledge in data and emphasizes the "high-level" application of data mining methods. The unifying goal of the KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. It does this by using data mining methods to extract what is deemed knowledge, Selecting database for data, Pre-processing, Transformation on data and Interpretation or Evaluation of patterns in knowledge.

CONCLUSION

This article focused on an important topic in the computer industry which helps in developing data mining models. One of the barriers of this article is the business understanding phase of the data mining process. Based on current research and polls conducted by website **KDNuggets,** CRISP-DM is the most widely used form of data-mining model in the data mining industries. Some of the drawbacks of this model is that it does not perform project management activities. The fact behind the success of CRISP-DM is that it is industry, tool, and application neutral. However, the major drawback is that the framework itself has not been updated on issues on working with new technologies, such as big data so there should be additional effort spent in the data understanding phase.

REFERENCES:

A. Schmidt, M. Atzmueller and M. Hollender, "Data Preparation for Big Data Analytics", *Advances in Business Information Systems and Analytics*, pp. 157-170, 2016. Available: 10.4018/978-1-5225-0293-7.ch010 [Accessed 3 March 2019].

S. Angée, S. Lozano-Argel, E. Montoya-Munera, J. Ospina-Arango and M. Tabares-Betancur, "Towards an Improved ASUM-DM Process Methodology for Cross-Disciplinary

Multi-organization Big Data & Analytics Projects", *Communications in Computer and Information Science*, pp. 613-624, 2018. Available: 10.1007/978-3-319-95204-8_51 [Accessed 3 March 2019].

"What main methodology are you using for your analytics, data mining, or data science projects? Poll", *Kdnuggets.com*, 2019. [Online]. Available: https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html. [Accessed: 03- Mar- 2019].

J. Taylor, "Four Problems in Using CRISP-DM and How To Fix Them", *Kdnuggets.com*, 2019. [Online]. Available: https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html. [Accessed: 03- Mar- 2019].

"Data mining", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Data_mining. [Accessed: 03- Mar- 2019].

"SEMMA", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/SEMMA. [Accessed: 03- Mar- 2019].

J. Stirrup, "What's wrong with CRISP-DM, and is there an alternative?", *Jen Stirrup*, 2019. [Online]. Available: https://jenstirrup.com/2017/07/01/whats-wrong-with-crisp-dm-and-is-there-an-alternative/. [Accessed: 03- Mar- 2019].

"The Forgotten Step in CRISP-DM and ASUM-DM Methodologies", *https://sharing.luminis.eu*, 2019. [Online]. Available: https://sharing.luminis.eu/blog/the-forgotten-step-in-crisp-dm-and-asum-dm-methodologies/. [Accessed: 03- Mar- 2019].

"Two popular Data Analytics methodologies every data professional should know: TDSP & CRISP-DM | Packt Hub", *Packt Hub*, 2019. [Online]. Available: https://hub.packtpub.com/two-popular-data-analytics-methodologies-every-data-professional-should-know-tdsp-crisp-dm/. [Accessed: 03- Mar- 2019].

"Team Data Science Process Documentation", *Docs.microsoft.com*, 2019. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/. [Accessed: 03- Mar- 2019].

J. Brownlee, "What is Data Mining and KDD", *Machine Learning Mastery*, 2019. [Online]. Available: https://machinelearningmastery.com/what-is-data-mining-and-kdd/. [Accessed: 03- Mar- 2019].

PART B

SMARTPHONE DATA MINING

Smartphone Data Mining

Mayank Ketan Gandhi
Masters in Data Analytics. Dublin Buisness School
Dublin, Ireland
10394669@mydbs.ie

***Abstract*—At first, mobile phones appeared in 1979 and since then there has been a tremendous growth in the mobile phone technology. Initially, mobile phones were only used for communication purpose only, but this changed with the invention of new generation of mobile devices called as smartphones. The extensive and world-wide use of smartphones has provided us with large amounts of data. These smart phones can provide us with data about various aspects such as location of the user, behaviour of the user, communication, user activity, etc. To find discrete user patterns and behaviour in enormous smart phone is one of the biggest problems that must be solved in the field of mobile data mining. We will discuss the process of mobile data mining and its efficient techniques and applications.**

***K*eywords:-** Mobile Data mining, Smartphones

I. Introduction

Mobile Data Mining involves the generation of interesting patterns after analysing the data collected from mobile phones and extracting useful information from it. The goal of mobile data mining is to provide advanced techniques for the analysis and monitoring of critical data and patters from the large amount of data. The patters found in the data are then filtered by means of relevancy and accuracy of the pattern and the final result of the pattern data is given to the decision maker. The decision maker can make use of this pattern data for his competitive advantage. Mobile data mining is an important field of research because it provides support for decision makers in order to make decisions relating to mobile users. The mobile data mining field may include several application scenarios in which a mobile device can play the role of data producer, data analyser, client of remote data miners, or a combination of them. The mobile phone has becoming an important device for providing information anytime anywhere.

II. The Data Mining Process

The smartphones are becoming an important device for providing information anytime anywhere. Acquiring data from these devices and processing it is not an easy task. There are many ways in which data from mobile can be mined and it undergoes various steps. These steps are as follows:

A. Collecting Data

The first process in Data Mining involves collection of data. There is various amounts of smartphone data that can be gathered like mobile application data, sensor data, location data, tracking of web usage data, etc. These data are periodically analysed.

B. ETL Process

The collected data then undergoes ETL (Extract, transform and load) processing. The mail goal is to extract the data, then transform into a clean format by removing errors and ambiguities from the data. The data is refined according to the business goals and needs. The final step involves loading this refined data to a data warehouse where all the data mining techniques are applied.

C. Mobile Data Mining

Once the data is loaded into the Data Warehouse, it is ready for various mining activities to be performed. Some of the applications of this analyzed data are discussed in detail.

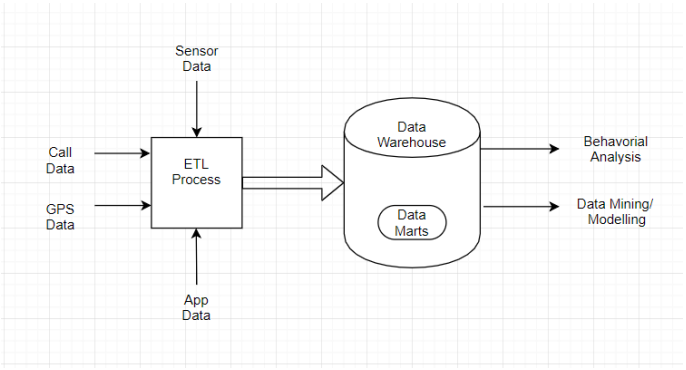

Fig. 1. Mobile Data Mining Process

III. Applications of Smartphone Data Mining

A. Mining Smart Phone Data for Personality Studies

Mobile phones have increasingly become an indispensable part of our daily lives. Considering the rapid growth of mobile phones, studying the psychological, social, and economic implications of smartphone usage has gained an increased importance. We can evaluate the relationship between the extracted behavioural characteristics that we get from rich smartphone data and self-reported Big-Five personality traits such as extraversion, agreeableness, conscientiousness, emotional stability and openness to experience. Since smartphones are programmable, they enable the development of data collection tools to record various behavioural aspects of the user, ranging from how the device is used across different contexts like analysing spatial and social dimensions of the of the user through sources such as GPS, call logs, and Bluetooth. The

tremendous growth of social media and its usage in smartphones mediate social interactions, social media could reflect an individual's personality. The ability to find patterns between personality and behavioural aspects derived through contextual data collected by mobile phones could lead to designing and applying data mining methods to classify users into personality types. For instance, could individuals who are introverts will be keen to using web-based communication channels could be verified with such models. While quantitative data analysis methods are suitable for highlighting statistical regularities, qualitative techniques are likely to be needed in order to obtain more insights into the reasons for individuals with a certain personality profile behaving in a given way. Support Vector Machines (SVM) Classifier or any classification algorithms can be used to classify user personality traits.

B. Smart Phone Based Data Mining for Human Activity Recognition

Smartphones, which are a new generation of mobile phones, are equipped with many powerful features including multitasking and a variety of sensors, in addition to the basic telephony. The integration of these mobile devices in our daily life is growing rapidly, and it is envisaged that such devices can seamlessly monitor and keep track of our activities, learn from them and assist us in making decisions. However, currently, though there is good capacity for collecting the data with such smart devices, there is limited capability in terms of automatic decision support capability and making sense out of this large data repository. User activities can be tracked with the help of sensor data, GPS data and application data from the mobile phones. Algorithm Used: Random Forests, Ensemble methods can be used for this kind of projects.

C. Mining Trajectory Data from Mobile

The Location-based services (LBSs) and its applications are a fast growing field because of the common use of GPS receivers and location sensing technology embedded in mobile devices. The Omnipresent GPS enabled mobile devices generate a huge amount of trajectory data, that are just the sequences of time-ordered locations of mobile devices. There has been a lot of work on collecting, storing, indexing and querying trajectories of mobile devices. Clustering trajectories of these mobile devices has a wide range of LBS applications. Application of mining trajectory data are, vehicular ad hoc network (VANET), traffic monitoring, transportation planning and location-based advertising.

D. Using Data Mining for Mobile Communication Clustering and Characterization

Telecommunication companies use data mining algorithms to analyse and profile their customers based on their communication behaviour. This analysis can be elaborated on call data available to any organization that uses any type of communication technologies (e.g. mobile, PBX, VoIP, teleconference). Data mining process is used in a large variety of activities in business, science and engineering to make better business decisions and strategies, by discovering patterns and relationships in the database. Most of the research in this domain shows that in the telecommunication industry, the data mining techniques can be used with success in application like market management, fraud detection and customer profiling. Furthermore, the problem of revealing users calling network profile and statistical calling information based on their calls it is essential in various applications, like customers churn management. Although the main beneficiaries of these studies are telecom service providers, companies working in other industries can benefit from analysing their communication patterns. Companies consuming telephony services have access to similar call detail data as their service providers. With this data we can investigate how clustering algorithms can emphasize users' communication patterns and identify users' communication behaviour.

IV. Conclusion

Mobile data mining is a fast-growing area of data mining which gives importance to the mobile phone user. The data collected from the mobile phones give invaluable knowledge representing various aspects of the user thus enabling the vendor to customize mobile usage as per the user needs. Despite the customization provided to the user, the privacy of the user is compromised. We have discussed areas where Mobile Data Mining is applicable and have demonstrated ways in which data can be mined in those areas.

References

[1] J. Goh and D. Taniar, "An Efficient Mobile Data Mining Model", Parallel and Distributed Processing and Applications, pp. 54-58, 2004. Available: 10.1007/978-3-540-30566-8-10 [Accessed 3 March 2019].

[2] Binh Han, Ling Liu and E. Omiecinski, "Road-Network Aware Trajectory Clustering: Integrating Locality, Flow, and Density", IEEE Transactions on Mobile Computing, vol. 14, no. 2, pp. 416-429, 2015. Available: 10.1109/tmc.2013.119.

[3] G. Chetty, M. White and F. Akther, "Smart Phone Based Data Mining for Human Activity Recognition", Procedia Computer Science, vol. 46, pp. 1181-1187, 2015. Available: 10.1016/j.procs.2015.01.031 [Accessed 3 March 2019].

[4] A. Bascacov, C. Cernazanu and M. Marcu, "Using data mining for mobile communication clustering and characterization", 2013 IEEE 8th International Symposium on Applied Computational Intelligence and Informatics (SACI), 2013. Available: 10.1109/saci.2013.6609004 [Accessed 3 March 2019].

[5] G. Chittaranjan, J. Blom and D. Gatica-Perez, "Mining large-scale smartphone data for personality studies", Personal and Ubiquitous Computing, vol. 17, no. 3, pp. 433-450, 2011. Available: 10.1007/s00779-011-0490-1 [Accessed 3 March 2019].

[6] "Data mining", En.wikipedia.org, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Data-mining. [Accessed: 03- Mar- 2019].

[7] S. AV, "Data Mining and Mobile Computing", Journal of Information Technology and Software Engineering, vol. 03, no. 01, 2013. Available: 10.4172/2165-7866.s6-e001 [Accessed 3 March 2019].