Data Mining Coursework Part A

Part 1

Data Mining, also known as 'Knowledge Discovery in Databases' is the process of discovering, analysing and using information gleaned from large datasets. This useful knowledge is usually patterns and relationships within the dataset. To do this, methods from a wide array of disciplines – from statistics to machine learning – are used. Generally speaking, data mining aims to find information about a group, rather than individuals.[1]

Data Mining's purpose is to manage and analyse big data. This can be further broken down into two ideas:

- Extracting information from the data
- And then using it to inform decision making. [2]

We can extract information in a few ways. We can use decision rules or decision trees, where we lay specific rules or conditions, which determine how the information is classified. We can also use equations to classify information, by calculating a value. Alternatively we can look at clusters, where the data is partitioned into groups based on how similar their data is.

We can use this information to inform decision making in several ways as well. We can use it in classification, where the data could be used to diagnose illness or machine faults, or in forecasting and prediction, where trends could be forecast ahead of time. We can also look for anomalous data, which could be used to identify frauds, for example.

Many different disciplines combine to make data mining effective, for example machine learning and image processing. Domain knowledge and image processing identifies key features and extracts the key parameter values, and then machine learning is used to work out the pattern.[2]

Figure 1[3] shows the data mining process. It involves collecting and preparing the data, doing the data mining, then evaluating and presenting any knowledge gathered. Depending on the results of the data mining, in the pattern evaluation stage (where the pattern outputted by the data mining tools are evaluated to see if it's good or not) the process may go back to the selection and transformation stage (where you select the data you want to process and transform it into a format that data mining tools can use) in order to try finding different, more useful patterns.

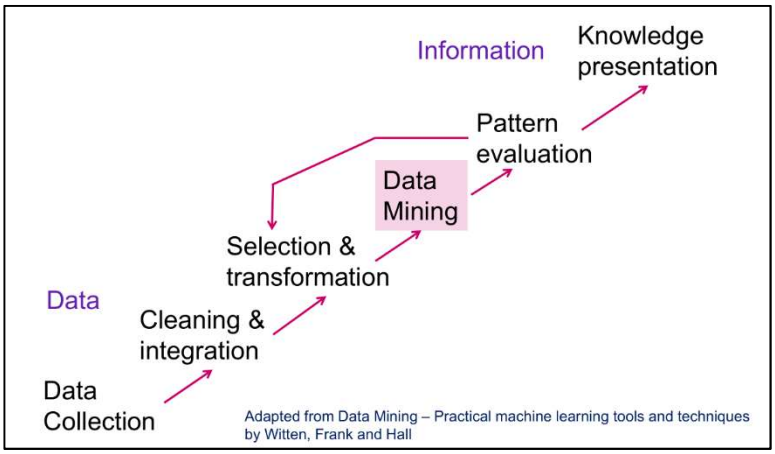

Figure 1 - Data Mining Process (accessed from Data Mining: Knowledge Discovery Process, 17 Mar 2020) [3]

In general, a good pattern is one that can be understood, has a valid outcome with a low level of uncertainty and, should be potentially useful in some way, which depends on how it will be used.

Decision trees are where each node is a decision point, and each leaf is a rule, as in Figure 2.

Rule Induction is the technique of generalising rules from specific cases. For example, after seeing 5 swans and they are all white, we might induce that all swans are white. These rules are unproven and uncertain. This can be used with decision trees to induce rules. [4]

Linear Regression is the technique of modelling the data to fit a line, and then using the equation of that line to help inform us of the data's characteristics.

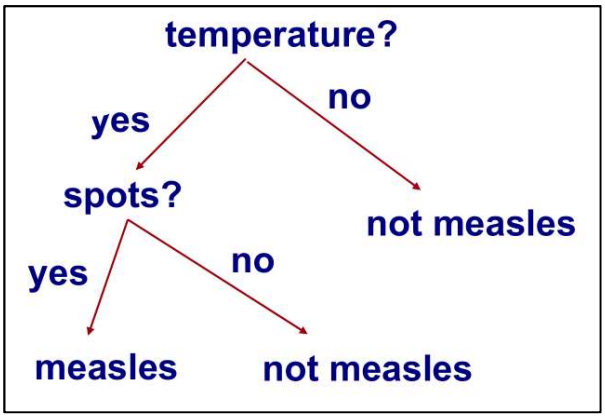

Figure 2 - Decision Tree (accessed from Data Mining: Techniques – Decision Tree and Rule Induction[4]

Instance-based classification is when the data set is searched for the instance that most closely resembles the new instance. This is a type of classification that is based on how similar different data are. Clustering is a very similar process where groups of close data points on a graph are grouped together.

The probability technique asks, 'what's the probability of the class, given an instance', which looks at the instance and its attributes and judges which class the instance is most likely to be in.

Part 2

There are many ways in which data mining can be applied to the healthcare sector. So much so that Patel and Patel (2016) [5] said that there are 'infinite applications' for data mining in their survey of data mining techniques that are used in the healthcare domain. They said that these techniques were able to 'discover hidden patterns and relationships' in medical data, which can be used in analysis, decision making, and to predict different kinds of diseases.

In 2011, Koh et al. [6] explored the applications of data mining within the healthcare sector, saying that data mining techniques allow the transformation of the large, complex amounts of data generated by healthcare transactions into useful information for decision making. This is emphasized by Srinivas et al. (2010) [7], who adds that 'there is a wealth of data available […]. However, there is a lack of effective analysis tools'. Data mining techniques are key when developing these analysis tools.

One example of this being done can be seen in Leary et al. (2020) [8], where she looks at the unused data of incident reporting systems (systems that manage the safety of patients in healthcare), which 'are commonly deployed in healthcare, but [the] resulting datasets are largely warehoused'. In her study, Leary identifies how intelligence from these systems can be used to identify patterns that can be used to improve care. As an example, she suggests using relationships found between staffing, nature of incidents and rates of reporting to inform decisions and review staffing levels. When data mining techniques are applied to data, it provides information that can be used at an organisational level.

Within healthcare data mining, it's important to detect anomalies, given we want to reduce our false-negative rate to 0%, and ideally reduce our false-positive rate as well. It is difficult to find an algorithm that can achieve both, but data mining techniques used alongside machine learning (specifically supervised learning) can achieve something that could optimise both. This is something that Ukil et al. (2016) [9] has explored, where they also note that there is a vast amount of

unstructured data (like clinical notes, journals), which can also be mined with anomaly detection in order to 'unearth huge information' to inform decision-making.

Data Mining can also be used to help inform the logistics of the healthcare sector. Teichmann et al. (2010) [10] explores how data mining can be used to predict the probability of whether or not patients will be re-admitted. This was to help prepare and mitigate rising admissions to the NHS. He explains how clinical data can be transformed into 'well-formed data sets' and then certain data mining rules can be applied – these rules were made using 'statistical characteristics of the data'. The tools proposed in this study could be used to reduce manual work and the predictive power of the dataset would hopefully increase.

Data Mining can also look at the data of each patient's length of stay, in order to inform prediction on other patients' length of stay. This has been looked at by Alahmar et al. (2018)[11] where he looks specifically at patients with diabetes.

Data Mining has also been used alongside image-based machine learning to help identify local dose-response relationships, as seen in Beasley et al. (2017) [12]. We can do a similar thing in tumour phenotyping as well, where we use a computational image-based model to increase the number and variance of the training set, thereby improving the quality of tumour phenotype predictions, which influence diagnosis and treatment of tumours. This can be seen in a patent by Comaniciu et al. (2017) [13].

Chaurasia et al. (2014) [14] also investigates how data mining can help detect tumours. They test different classification methods to try and develop accurate prediction models. They note that a 'challenge in data mining and machine learning areas is to build precise […] classifiers for Medical applications'.

Data mining can be used in similar ways to identify and predict various other conditions and can be used to influence diagnosis and treatment. Using data mining analytics, Hogg et al. (2018) [15] tries to 'characterise the multiple sclerosis (MS) prodrome'. Some symptoms of MS are believed to manifest some years in advance of diagnosis, so identifying characteristics that might be used to predict the development of MS would greatly impact diagnosis, as well as contributing to a better understanding of the MS prodrome.

There are many elements that affect the development of heart disease, and various data mining techniques, as identified by Thomas et al. (2016) [16], can be used to predict the risk of heart disease. The study finds that 'various data mining techniques and classifiers […] are used for efficient and efficacious heart disease'. Data mining techniques, as was used with the MS prodrome, could also be used to help identify further patterns and further risk factors for heart disease.

Beyond using data mining for predicting conditions and analysing data to improve care and gain knowledge for decision making, data mining can also be used to detect fraud. In their study, Bauder et al. (2017) [17] evaluates how different techniques can be applied to the large amount of healthcare data to detect fraud. They state that using these techniques 'has the potential to greatly reduce healthcare costs through a more robust detection of fraud'. They do note that 'audit data can be difficult to obtain, limiting the usefulness of supervised learning.

Joudaki et al. (2016) [18] explores this further, looking at fraud and also abuse of drug prescription claims. They used a data mining approach and identified 13 indicators.

Many studies note that there is a vast amount of data generated by the healthcare sector, but a lack of effective analysis being done. Research is being done, and various tools are being developed, though it will still take some years before all the data generated is effectively analysed.

Data mining systems can be used to analyse this large amount of data to inform decision making. Due to the diverse amount of data available, the types of decisions that can be affected are also very diverse. For example, data from incident reporting systems can improve patient's care, while we can use data about a patient's hospital stay to inform us about their risk of readmission and to predict other patients' length of stays.

Data mining systems work well with various AI systems, particularly image-based machine learning. This combination can be used together to build systems that predict conditions, such as predicting tumour phenotypes, heart disease and the MS prodrome, among others. Not only does this inform diagnoses and treatments, but these can also contribute to a better understanding of the conditions, especially if the system identifies a parameter that was previously thought to be unrelated.

Furthermore, data mining can also be applied with regards to fraud and abuse detection, which can lead to reduced costs and a more reliable fraud and abuse detection system.

The potential is massive for using data mining in the healthcare sector. There are many datasets that can be applied to a huge number of different areas. Almost any area can use data mining techniques to inform decisions, predict conditions and also improve the day-to-day running and logistics of the healthcare sector.

Part 3

In order to introduce data mining to our company, I've considered each stage of the data mining process. I'll be using Figure 1 from Part 1 of this report as a guide.[3]

First of all, it is important to identify the problem we want to solve and maybe establish some goals. For example, we could aim to create a system that looks at the prediction and diagnoses of motor-neuron disease, or ALS. Using this goal as a baseline, we can then start collecting relevant data with which we can build our dataset.

'Data Collection' is the first stage of the 'Knowledge Discovery' process. There is a large variety of the types of data we could collect, from existing structured data such as databases to unstructured data such as those in clinical notes. It's important to take some time to properly identify all the data that we want. There could be some issues gaining access to some data. Data might need to be shared with us from independent sources, such as hospitals, and this could pose several issues.

Healthcare data can be sensitive and includes a lot of personal data. In order to follow legislation like GDPR, and also to comply with the independent source's policies on confidentiality and privacy, we would need to have certain measures in place to properly protect and safeguard any personal information. This would enable us to assure those sources that the data is safe. We would also need to build a network of connections in order to gather data from a wide array of sources, in order to increase the generalisability of any systems we build.

Data cleaning and integration is the next step, where we remove noise, inconsistencies and duplicates and deal with missing values. This is also where we combine multiple sources together, which could pose an issue or two. If we are look at a variety of sources, for example a pile of completed forms and a database, combining these data could take some work to standardize and combine. It could be worth focusing on one type of data for each data mining system instead. Also,

when accepting the same type of data from a number of sources like several different hospitals, there might not a standardized method of storing the data, which means the data from each source could be in slightly different formats, which would take longer to combine as they would have to be integrated properly.

The next stage is selection and transformation. This is where we look closely at our data, identify what is relevant and then transform the data into a format that the data mining system can use. This should be straightforward if it's been cleaned and integrated properly. It may take some time to select the relevant data if it's an obscure area with not much research or no existing data mining systems and may need an expert medical opinion if involving a medical condition.

Data mining is the next stage, where we use our chosen techniques to try and identify patterns in the data. We can try a variety of data mining techniques, some of which are laid out in Part 1 of this report.

After the data mining system has run through the dataset, we need to evaluate the pattern that has been found. We need to check that the pattern is understandable, potentially useful and also a valid outcome with an acceptable level of clarity. The 'acceptable level' mentioned here could be quite a high level, depending on our goal. If we are looking at anything involving prediction or diagnoses of any illnesses, as the outcome could have significant impact, we want to have a high level of confidence and clarity in our outcome. We also want to pay attention to the possibility of anomalies. False-positives and false-negatives would have ethical issues – if our result contributed to a diagnosis that was a false-positive, we would deal unnecessary psychological harm, worry and perhaps trauma. A false-negative in the same situation would be worse, as it would lead to a missed or delayed diagnosis, leading to potentially physical harm.

If the pattern isn't interesting or useful in any way, we can go back and retry the process from the 'selection and transformation' stage, maybe selecting different data or using different data mining techniques.

After finding a pattern that is acceptable, we then present the knowledge we've gained. This can range from informing decisions and diagnoses to predicting the likelihood of developing a condition, or maybe just contributing to the scientific understanding about a condition. Depending on how impactful a decision or diagnoses is, we should be careful to make sure our outcome is accurate.

It could be dangerous to rely on the outcome fully, as the accuracy will almost never be 100%, and there will likely always be a margin of error. Therefore, it would likely be sensible to not over-rely on the outcome of these data mining systems, but rather using the outcomes to inform any decisions we make, rather than using it to directly decide.

References

1. Clifton C 2010, *Definition of Data Mining*, Encyclopedia Brittanica, viewed 17 Mar 2020 (https://www.britannica.com/technology/data-mining)
2. Derakhshan, P 2020, *Data Mining: Purpose and Applications*, lecture notes, Data Mining 19COC131, Loughborough University, delivered 10th Feb 2020.
3. Derakhshan, P 2020, *Data Mining: Knowledge Discovery Process*, lecture notes, Data Mining 19COC131, Loughborough University, delivered 17th Feb 2020.
4. Derakhshan, P 2020, *Data Mining: Decision Tree and Rule Induction*, lecture notes, Data Mining 19COC131, Loughborough University

5. Patel S, Patel H 2016, 'Survey Of Data Mining Techniques Used In Healthcare Domain', *International Journal of Information Sciences and Techniques (IJIST), Vol. 6, No. 1/2, pp. 53-60.*
6. Koh H, Tan G 2011, 'Data Mining Applications in Healthcare', *Journal of Healthcare Information Management*, Vol.19, No.2, pp. 64-72
7. Srinivas K, Rani B, Govrdhan A 2010, 'Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks', *(IJCSE) International Journal on Computer Science and Engineering,* Vol. 2, No. 2, pp. 250-255
8. Leary A, Cook R, Jones S, Radford M, Smith J, Gough M, Punshon G 2020, 'Using knowledge discovery through data mining to gain intelligence from routinely collected incident reporting in an acute English hospital', *International Journal of Health Care Quality Assurance*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/IJHCQA-08-2018-0209
9. Ukil A, Bandyoapdhyay S, Puri C, Pal A 2016, 'IoT Healthcare Analytics: The Importance of Anomaly Detection', *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, Crans-Montana, pp. 994-997.
10. Teichmann E, Demir E, Chaussalet T 2010, 'Data preparation for clinical data mining to identify patients at risk of readmission', *2010 IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS)*, Perth, WA, pp. 184-189.
11. Alahmar A, Mohammed E, Benlamri R 2018, 'Application of Data Mining Techniques to Predict the Length of Stay of Hospitalized Patients with Diabetes,' *2018 4th International Conference on Big Data Innovations and Applications (Innovate-Data)*, Barcelona, pp. 38-43.
12. Beasley W, Thor M, McWilliam A, Green A, Mackay R, Slevin N, Olsson C, Pettersson N, Finizia C, Deasy J, Van Herk M 2017, 'Image-Based Data Mining for Identifying Regions Exhibiting a Dose-Response Relationship with Radiation-Induced Trismus', *International Journal of Radiation Oncology*, Vol. 99, No. 2, Supplement, Page S165.
13. Comaniciu D, Kamen A, Liu D, Mailhe B, Mansi T 2016, Image-based tumor phenotyping with machine learning from synthetic data, US10282588B2
14. Chaurasia V, Pal S 2014, 'A Novel Approach for Breast Cancer Detection Using Data Mining Techniques', *International Journal of Innovative Research in Computer and Communication Engineering,* Vol. 2, No. 1
15. Högg T, Wijnands J, Kingwell E, Zhu F, Lu X, Evans C, Fisk J, Ann-Marrie R, Zhao Y, Tremlett H 2018, 'Mining healthcare data for markers of the multiple sclerosis prodrome', *Multiple Sclerosis and Related Disorders*, Vol 25, pp. 232-240
16. Thomas J, Princy RT 2016, 'Human heart disease prediction system using data mining techniques,' 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Nagercoil, pp. 1-5
17. Bauder R, Khoshgoftaar TM, Seliya N 2017, 'A survey on the state of healthcare upcoding fraud analysis and detection', *Health Services and Outcomes Research Methodology, Vol.* 17, pp. 31–55
18. Joudaki H, Rashidian A, Minaei-Bidgoli B, Mahmoodi M, Geraili B, Nasiri M, Arab M 2016, 'Improving Fraud and Abuse Detection in General Physician Claims: A Data Mining Study', *Int J Health Policy Manag,* Vol 5, No. 3, pp. 165-172