THE IEEE

Intelligent Informatics

BULLETIN

IEEE Computer Society
Technical Committee
on Intelligent Informatics

August 2018 Vol. 19 No. 1 (ISSN 1727-5997)

Feature Articles

Selected PhD Thesis Abstracts

On-line version: http://www.comp.hkbu.edu.hk/~iib (ISSN 1727-6004)

IEEE Computer Society Technical Committee on Intelligent Informatics (TCII)

Executive Committee of the TCII:

The Technical Committee on Intelligent Informatics (TCII) of the IEEE Computer Society deals with tools and systems using biologically and linguistically motivated computational paradigms such as artificial neural networks, fuzzy logic, evolutionary optimization, rough sets, data mining, Web intelligence, intelligent agent technology, parallel and distributed information processing, and virtual reality. If you are a member of the IEEE Computer Society, you may join the TCII without cost at http://computer.org/tcsignup/.

The IEEE Intelligent Informatics Bulletin

Aims and Scope

The IEEE Intelligent Informatics Bulletin is the official publication of the Technical Committee on Intelligent Informatics (TCII) of the IEEE Computer Society, which is published twice a year in both hardcopies and electronic copies. The contents of the Bulletin include (but may not be limited to):

1) Letters and Communications of the TCII Executive Committee

2) Feature Articles

3) R&D Profiles (R&D organizations, interview profile on individuals, and projects etc.)

4) Selected PhD Thesis Abstracts

5) Book Reviews

6) News, Reports, and Announcements (TCII sponsored or important/related activities)

Materials suitable for publication at the IEEE Intelligent Informatics Bulletin should be sent directly to the Associate Editors of respective sections.

Technical or survey articles are subject to peer reviews, and their scope may include the theories, methods, tools, techniques, systems, and experiences for/in developing and applying biologically and linguistically motivated computational paradigms, such as artificial neural networks, fuzzy logic, evolutionary optimization, rough sets, and self-organization in the research and application domains, such as data mining, Web intelligence, intelligent agent technology, parallel and distributed information processing, and virtual reality.

Editorial Board

Publisher: *The IEEE Computer Society Technical Committee on Intelligent Informatics*
Address: *Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (Attention: Dr. William K. Cheung; Email:william@comp.hkbu.edu.hk)*
ISSN Number: *1727-5997(printed)1727-6004(on-line)*
Abstracting and Indexing: *All the published articles will be submitted to the following on-line search engines and bibliographies databases for indexing—Google(www.google.com), The ResearchIndex(citeseer.nj.nec.com), The Collection of Computer Science Bibliographies (liinwww.ira.uka.de/bibliography/index.html), and **DBLP** Computer Science Bibliography (www.informatik.uni-trier.de/~ley/db/index.html).*

Interpretable Machine Learning in Healthcare

Muhammad Aurangzeb Ahmad, Carly Eckert, Ankur Teredesai, and Greg McKelvey

***Abstract*—The drive towards greater penetration of machine learning in healthcare is being accompanied by increased calls for machine learning and AI based systems to be regulated and held accountable in healthcare. Interpretable machine learning models can be instrumental in holding machine learning systems accountable. Healthcare offers unique challenges for machine learning where the demands for explainability, model fidelity and performance in general are much higher as compared to most other domains. In this paper we review the notion of interpretability within the context of healthcare, the various nuances associated with it, challenges related to interpretability which are unique to healthcare and the future of interpretability in healthcare.**

***Index Terms*—Interpretable Machine Learning, Machine Learning in Healthcare, Health Informatics**

I. Introduction

WHILE the use of machine learning and artificial intelligence in medicine has its roots in the earliest days of the field [1], it is only in recent years that there has been a push towards the recognition of the need to have healthcare solutions powered by machine learning. This has led researchers to suggest that it is only a matter of time before machine learning will be ubiquitous in healthcare [22]. Despite the recognition of the value of machine learning (ML) in healthcare, impediments to further adoption remain. One pivotal impediment relates to the *black box* nature, or opacity, of many machine learning algorithms. Especially in critical use cases that include clinical decision making, there is some hesitation in the deployment of such models because the cost of model misclassification is potentially high [21]. Healthcare abounds with possible "high stakes" applications of ML algorithms: predicting patient risk of sepsis (a potentially life threatening response to infection), predicting a patient's likelihood of readmission to the hospital, and predicting the need for end of life care, just to name a few. Interpretable ML thus allows the end user to interrogate, understand, debug and even improve the machine learning system. There is much opportunity and demand for interpretable ML models in such situations. Interpretable ML models allow end users to evaluate the model, ideally before an action is taken by the end user, such as the clinician. By explaining the reasoning behind predictions, interpretable machine learning systems give users reasons to accept or reject predictions and recommendations.

Audits of machine learning systems in domains like healthcare and the criminal justice system reveal that the decisions and recommendations of machine learning systems may be biased [4]. Thus, interpretability is needed to ensure that such systems are free from bias and fair in scoring different ethnic and social groups [12]. Lastly, machine learning systems are already making decisions and recommendations for tens of millions of people around the world (i.e. Netflix, Alibaba, Amazon). These predictive algorithms are having disruptive effects on society [32] and resulting in unforeseen consequences [12] like deskilling of physicians. While the application of machine learning methods to healthcare problems is inevitable given that complexity of analyzing massive amounts of data, the need to standardize the expectation for interpretable ML in this domain is critical.

Historically, there has been a trade-off between interpretable machine learning models and performance (precision, recall, F-Score, AUC, etc.) of the prediction models [8]. That is, more interpretable models like regression models and decision trees often perform less well on many prediction tasks compared to less interpretable models like gradient boosting, deep learning models, and others. Researchers and scientists have had to balance the desire for the most highly performing model to that which is adequately interpretable. In the last few years, researchers have proposed new models which exhibit high performance as well as interpretability e.g., GA2M [5], rule-based models like SLIM[30], falling rule lists[31], and model distillation [27]. However, the utility of these models in healthcare has not been convincingly demonstrated due to the rarity of their application.

The lack of interpretability in ML models can potentially have adverse or even life threatening consequences. Consider a scenario where the insights from a black box models are used for operationalizating without the recognition that the predictive model is not prescriptive in nature. As an example, consider Caruana *et al.* [5] work on building classifiers for labeling pneumonia patients as high or low risk for in-hospital mortality. A neural network, essentially a black box in terms of interpretability, proved to be the best classifier for this problem. Investigation of this problem with regression models revealed that one of the top predictors was *patient history of asthma*, a chronic pulmonary disease. The model was predicting that given asthma, a patient had a lower risk of in-hospital death when admitted for pneumonia. In fact, the opposite is true - patients with asthma are at higher risk for serious complications and sequelae, including death, from an infectious pulmonary disease like pneumonia. The asthma patients were, in fact, provided more timely care of a higher acuity than their counterparts without asthma, thereby incurring a survival advantage. Similarly leakage from data can misinform models or artificially inflate performance during testing [14], however explanations can be used to interrogate and rectify models when such problems surface.

While there is a call to apply interpretable ML models to a large number of domains, healthcare is particularly challenging due to medicolegal and ethical requirements, laws, and regulations, as well as the very real caution that must be

The authors are from KenSci Inc. Corresponding Author e-mail: (muhammad@kensci.com).

employed when venturing into this domain. There are ethical, legal and regulatory challenges that are unique to healthcare given that healthcare decisions can have an immediate effect on the wellbeing or even the life of a person. Regulations like the European Union's General Data Protection Regulation (GDPR) require organizations which use patient data for predictions and recommendations to provide *on demand* explanations [28]. The inability to provide such explanations on demand may result in large penalties for the organizations involved. Thus, there are monetary as well as regulatory and safety incentives associated with interpretable ML models.

Interpretability of ML models is applicable across all types of ML: supervised learning [17], unsupervised learning [6] and reinforcement learning [15]. In this paper, we limit the scope of the discussion to interpretability in supervised learning models as this covers the majority of the ML systems deployed in healthcare settings [18]. The remainder of the paper is organized as follows: First, we define interpretability in machine learning, we provide an overview of the need for interpretability in machine learning models in healthcare, and we discuss use cases where interpretability is less critical. We conclude this paper with a brief survey of interpretable ML models and challenges related to interpretability unique to healthcare.

II. What is Interpretability?

While there is general consensus regarding the need for interpretability in machine learning models, there is much less agreement about what constitutes interpretability [17]. To this end, researchers have tried to elucidate the numerous notions and definitions of interpretability [17],[8]. Interpretability has been defined in terms of model transparency [17], model fidelity [17], model trust [17], [8], [9], and model comprehension [9], among other characteristics. Many of the notions of interpretability have been developed in the context of computing systems and mostly ignore the literature on interpretability that comes from the social sciences or psychology [19]. Thus, one common objection to these definitions of interpretability is that it does not put enough emphasis on the user of interpretable machine learning systems [16]. This results in a situation where the models and explanations produced do not facilitate the needs of the end users [19].

A primary sentiment of interpretability is the fidelity of the model and its explanation i.e., the machine learning model should give an explanation of why it is making a prediction or giving a recommendation. This is often referred to as a key component of "user trust" [25]. In some machine learning models like decision trees [24], regression models [33], and context explanation networks [3] the explanation itself is part of the model. In contrast, for models such as neural networks, support vector machines, and random forests that do not have explanations as part of their predictions it is possible to extract explanations from models that are applied post-hoc, such as locally interpretable model explanations (LIME) [25] and Shapley Values [26]. LIME constructs explanations by creating a local model, like a regression model, for the instance for which an explanation is required. The data for the local model is generated by perturbing the instance of interest, observing the change in labels and using it to train a new model. Shapley values, on the other hand, take a game theoretical perspective to determine the relative contribution of variables to the predictions by considering all possible combinations of variables as cooperating and competing coalitions to maximize payoff, defined in terms of the prediction [26].

Many definitions of interpretability include transparency of the components and algorithms, the use of comprehensible features in model building, and intelligible applications of parameters and hyperparameters. Based on the work of Lipton *et al.* [17], interpretability can be described in terms of transparency of the machine learning system *i.e.,* the algorithm, features, parameters and the resultant model should be comprehensible by the end user. At the feature level, the semantics of the features should be understandable. Thus, a patient's age is readily interpretable as compared to a highly engineered feature (the third derivative of a function that incorporates age, social status and gender, for example). At the model level, a deep learning model is less interpretable compared to a logistic regression model. An exception to this rule is when the deep learning model utilizes intuitive features as inputs and the regression model utilizes highly engineered features, then the deep learning model may in fact be more interpretable. Lastly, we consider interpretability in terms of the model parameters and hyperparameters. From this perspective, the number of nodes and the depth of the neural network is not interpretable but the number of support vectors for a linear kernel is much more interpretable [17].

Interpretability may also mean different things for different people and in different use cases. Consider regression models. For a statistician or a machine learning expert the following equation for linear regression is quite interpretable:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, ..., n \tag{1}$$

Those familiar with the field, can easily identify the relative weights of the parameter coefficients and abstract meaning from the derived equation. However, most non statisticians, including some clinicians, may not be able to interpret the meaning of this equation. For others, merely describing a model as "linear" may be sufficient. Conversely, a more advanced audience, knowing the error surface of the model may be needed to consider the model fully "interpretable".

For some predictive algorithms, however, the lack of interpretability may go deeper. Thus consider the following equation for updating weights in a deep learning network.

$$a_j^l = \sigma\Big(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\Big) \tag{2}$$

While the math is clear and interpretable, the equation does not help anyone understand how deep learning networks actually learn and generalize.

Finally, interpretability of machine learning models in healthcare is always context dependent, even to the level of the user role. The same machine learning model may require generating different explanations for different end users e.g., an explanation model for a risk of readmission prediction model to be consumed by a hospital discharge planner vs.

a physician may necessitate different explanations for the same risk score. This component of interpretability parallels the thought processes and available interventions of different personas in healthcare. For example, a discharge planner will often evaluate a patient's risk of readmission based on the components of that patient's situation that are under her purview - perhaps related to the patient's living situation, unreliable transportation, or need for a primary care physician. While the treating physician will need to be aware of these associated characteristics, she may be more likely to focus on the patient's cardiac risk and history of low compliance with medications that are associated with the patient's high risk of readmission. Context is critical when considering interpretability.

III. Interpretability Vs. Risk

While there are a number of reasons why interpretability of ML models is important, not all prediction problems in supervised machine learning predictions require explanations. Alternatives to explanations include domains where the system may have theoretical guarantees to always work or empirical guarantees of performance when the system has historically shown to have great performance e.g., deep learning applications radiology with superhuman performance[20]; or in work pioneered by Gulshan et al, the developed deep learning algorithm was able to detect diabetic retinopathy from retinal fundal photographs with extremely high sensitivity and specificity [10]. The exceptional performance supports the fact that this prediction does not require an explanation. However, findings such as this are quite rare. Another example where interpretability may not be prioritized is in the setting of emergency department (ED) crowding. For a hospital's ED, the number of patients expected to arrive at the ED in the next several hours can be a helpful prediction to anticipate ED staffing. In general, the nursing supervisor is not concerned with the reasons why they are seeing the expected number of patients (of course, there are exceptions) but only interested in the number of expected patients and the accuracy of the prediction. On the other hand, consider the case of predicting risk of mortality for patients. In this scenario, the imperative for supporting explanations for predictions may be great - as the risk score may drive critical care decisions. What these examples demonstrate is that the clinical context (also, how ”close” the algorithm is to the patient) associated with the application determines the need for explanation. The fidelity of the interpretable models also plays a role in determining the need for explanations. Models like LIME [25] produce explanations which may not correspond to how the predictive model actually works. LIME models are post-hoc explanations of model output, and in some ways, likely mimics the manner in which human beings explain their own decision making processes [17], this may be an admissible explanation where explanations are needed but the cost for the occasional false positives is not very high.

Consider Figure 1 which shows a continuum of potential risk predictions related to patient care. The arrow represents the increasing need for explanations along the continuum. Consider a model for cost prediction for a patient, the accuracy of the prediction may take precedence over explanation depending on the user role. However, as we move up the continuum to *Length of hospital stay* explanations may be helpful in decision making while tolerating a slight decrement in model performance. Thus, the specific use case is very important when considering which predictive and explanation models to choose. Certain use cases and domains require us to sacrifice performance for interpretability while in other cases, predictive performance may be the priority.

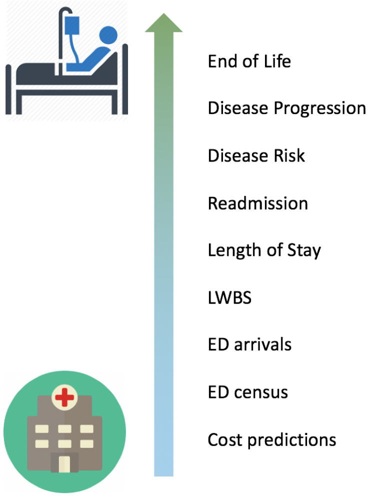

Fig. 1: Prediction Use Cases vs. Need for Interpretability (LWBS: left without being seen)

IV. The Challenge of Interpretability in Healthcare

The motivation for model explanations in healthcare is clear - in many cases both the end users and the critical nature of the prediction demand a certain transparency - both for user engagement and for patient safety. However, merely providing an explanation for an algorithm's prediction is insufficient. The manner in which interpretations are shared with the end users, incorporated into user workflows, and utilized must be carefully considered.

Healthcare workers are generally overwhelmed - by the number of patients they are required to see in a shift, by the amount of data generated by such patients, and the associated tasks required of them (data entry, electronic health record system requirements, as well as providing clinical care). Machine learning algorithms and their associated explanations, if not delivered correctly, will merely be one additional piece of data delivered to a harried healthcare professional. In order to be truly considered, ML output should be comprehensible to the intended user from a domain perspective and be applicable with respect to the intended use case.

A. User Centric Explanations

The participation of end users in the design of clinical machine learning tools is imperative - to better understand how the end users will utilize the output components - and

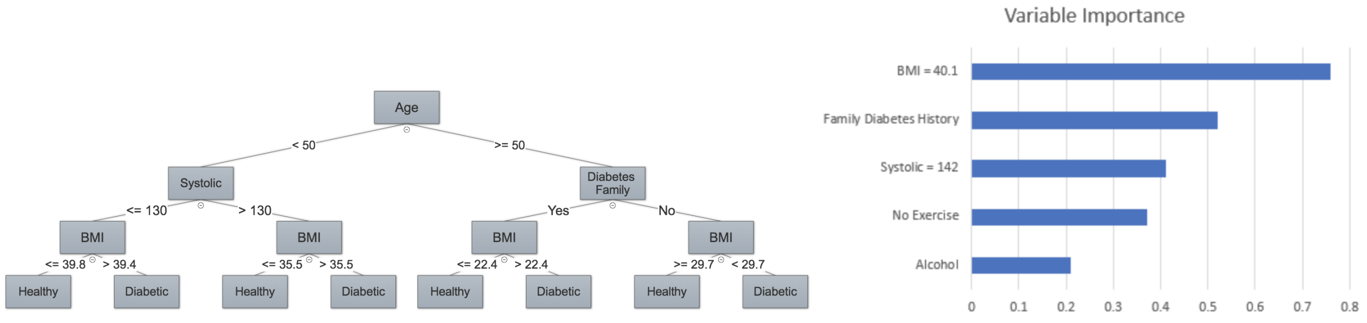

Fig. 2: Global vs. Local Models for Predicting Diabetes

also to educate end users to the nature of the prediction and explanation. According to Jeffrey *et al.* [13], even seasoned clinicians have difficulty interpreting risk scores and probability based estimates and end user input in the design of the expected output can drive participation. Moreso, understanding how end users interpret explanations, derived from the machine learning models, is imperative. Consider for example, the following output:

Patient risk of Readmission: 62, HIGH
Top Factors: Low albumin
Elevated heart rate in emergency department
History of heart failure

How will a healthcare provider interpret this resulting risk score and the associated explanation? Does the fact that the provider knows that these attributes are ”true” for this patient allow user trust in the model? Does the physician consider that by addressing these top factors - such as the patient’s low albumin- that the patient’s risk of readmission will be mitigated? It is important that the concepts of causality and association are emphasized and differentiated. Lipton [17] addresses the issue of algorithm explanation and the tendency to attribute causality to an explanation. He cautions against the conflation of these concepts but does remind us that the results of the explanation can instead inform future formal studies to investigate causal associations between the associated factor and the end point, i.e. readmission.

B. Performance vs. Transparency Trade-off

Earlier in this paper we described the trade-off between model performance and model transparency in healthcare algorithms. How is this trade-off determined? and by whom? Others have described the need to optimize models towards different performance metrics, and that AUC may not always be the metric to optimize. For example, when predicting end of life to determine when to refer patients to hospice, physicians may prefer to optimize for model precision, that is, to maximize the number of individuals who are correctly classified as likely to die by the algorithm. Similarly, the trade-off between performance and interpretability requires discussions with end users to understand the clinical and human risk associated with misclassification or with model opacity.

C. Balancing Requirements of Interpretability

As there is not a single unified conception of interpretability, there are multiple requirements for an ideal interpretable machine learning system, some of which may be at odds. Consider model soundness which refers to how close the model explanation is to how the model actually works. It may be the case that the model which results in the best performance and interpretability is a decision trees with depth 8 and 50 nodes. While the decision tree model is interpretable, the whole model is not comprehensible at the same time. Simultability is a characteristic of a model when it can be comprehensible in its entirety [17]. In this situation, it may be possible to make the decision tree more interpretable by pruning and then use that model for explanations. This may result in a loss in performance and also a loss in soundness, as the model now corresponds to a lesser degree regarding how predictions are being made.

Certain healthcare applications such as predicting disease progression may require explanations at the local, the cohort and the global level. For such applications, local explanations like LIME or Shapley Values may not suffice. One way to address the requirements of explanation scope is to first generate the local explanations first and to then generate global level explanations by aggregating these. The main drawback in such approaches is the large runtime required to generate explanations for individual instances. Another way to address this problem is to create distilled models like decision trees for generating global explanations and local models for explanations at the instance level.

Lastly, trust is one of the most important aspects of interpretability. Consider the case of deep learning models that have shown great predictive performance in a number of healthcare applications [20]. While it is possible to extract explanations from deep learning models, these explanations cannot be proven to be sound or complete [23]. Often the goal of explanations is to get parsimonious explanations which cannot be stated to have the correct explanations. Additionally always having parsimony as a goal may lead to incorrect models [7].

D. Assistive Intelligence

One common misconception about the application of machine learning in healthcare is that machine learning algorithms are intended to replace human practitioners in healthcare and medicine [11]. Healthcare delivery is an extremely complex, subtle, and intimate process that requires domain knowledge and intervention in every step of care. We believe that the human healthcare practitioner will remain integral to their role and that machine learning algorithms can assist and augment the provision of better care. Human performance parity [17] is also considered to be an important aspect of predictive systems that provide explanations i.e., the predictive system should be at least as good as the humans in the domain and at least make the same mistakes that the human is making. In certain use cases the opposite requirement may hold i.e., one may not care about parity with cases when humans are right but rather one cares more about cases where humans are bad at prediction but the machine learning system has superior performance. Such hybrid of human-machine learning systems can lead to truly assistive machine Learning in healthcare. Explanations from such systems could also be used to improve human performance, extract insights, gain new knowledge which may be used to generate hypothesis etc. The results from hypothesis derived from the data driven paradigm could in turn be used to push the frontiers of knowledge in healthcare and medicine by guiding theory [2].

V. Interpretable Models in Healthcare

Depending upon the scope of the problem, explanations from machine learning models can be divided into “global”, “cohort-specific” and “local” explanations. Global explanations refer to explanations that apply to the entire study population e.g., in the case of decision trees and regression models. Cohort-specific explanations are explanations that are focal to population sub-groups. Local explanations refer to explanations that are at the instance level i.e., explanations that are generated for individuals. Consider Figure 2 which illustrates the contrast between global vs. local models for predicting diabetes. The global model is a decision tree model that generalizes over the entire population, the cohort level model can also be a decision tree model which captures certain nuances of the sub-population of patients not captured by the global model and lastly the local model gives explanations at the level of instances. All three explanations may be equally valid depending upon the use case and how much soundness and generalizability is required by the application.

One way to distinguish models is by model composition. The predictive model and the explanation of the model can be the same as in the case of decision trees, GA2M etc. Alternatively they can be different e.g., a Gradient Boosting model is not really interpretable but it is possible to extract explanations via models like LIME, Shapley values, Tree Explainers etc. One scheme to create interpretable models is via model distillation where the main idea is to create interpretable models from non-interpretable models. Consider a feature set $X = x_1, x_2, x_3, ..., x_n$ with y_i is the class label being predicted. Suppose y'_i is the label that is predicted by a prediction model M_p which is non-interpretable *e.g.,* Deep Learning etc. An interpretable model *e.g.,* decision trees, regression models etc. which is created by the feature set X and the output y'_i as the label is referred to as a student model. While there are no theoretical guarantees for the performance of the student model but in practice, many student models have predictive power which is sufficiently high from an application perspective.

VI. Future of Interpretability in Healthcare

As machine learning increasingly penetrates healthcare, issues around accountability, fairness and transparency of machine learning systems in healthcare will become paramount. Most predictive machine learning systems in healthcare just provide predictions but in practice many use cases do require reasoning to convince medical practitioners to take feedback from such models. Thus there is a need to integrate interpretable models with predictions with the workflow of medical facilities. Most predictive models are not prescriptive or causal in nature. In many healthcare applications explanations are not sufficient and prescriptions or actionability. We foresee causal explanations to be the next frontier of machine learning research.

It should also be noted that while interpretability is an aspect of holding machine learning models accountable, it is not the only way to do. Researchers have also suggested that one way to audit machine learning systems it to analyze their outputs given that some models may be too complex for human comprehension [29] and auditing outputs for fairness and bias may be a better option. Also, many problems in healthcare are complex and simplifying them to point solutions with accompanying explanations may result in suboptimal outcomes. Thus consider the problem of optimizing risk of readmission to a hospital. Just optimizing predictions and actionability to reducing risk of readmission may in fact increase the average length of stay in hospitals for patients. This would be non-optimal solution and not in the best interest of the patient even though the original formulation of the machine learning problem is defined as such. Thus problem formulations for interpretable models should take such contexts and interdependencies into account.

There is also some debate around the use of post-hoc vs. ante-hoc models of prediction in the research community. Since explanations from post-hoc models do not correspond to how the model actually predicts, there is skepticism regarding the use of these models in scenarios which may require critical decision making. Current and future efforts in predictive models should also focus on ante-hoc explanation models like context explanation networks, falling rule lists, SLIM etc. Scalability of interpretable machine learning models is also an open area of research. Generating explanations for models like LIME and Shapley values can be computationally expensive. In case of LIME, a local model has to be created for each instance for which an explanation is required. In a scenario where there are hundreds of millions of instances for which prediction and explanations are required then this can be problematic from a scalability perspective. Shapley values computation requires computing the variable contribution by considering all possible combinations of variables in the data. For problems where the feature set has hundreds of variables, such computations can be very expensive. The problem of scalability thus exists with two of the most widely used interpretable machine learning models.

Lastly, evaluation of explanation models is an area which has not been explored in much detail. Consider the scenario in which multiple models with the same generalization error offer different explanations for the same instance or alternatively different model agnostic models are used to extract explanations and these model offer different explanations. In both these scenarios, the challenge is to figure out which explanations are the best. We propose that the concordance in explanations as well as how well the explanations align with what is already known in the domain will determine explanation model preference. However, the danger also exists that novel but correct explanations may be weeded out if concordance is the only criteria of choosing explanations.

VII. Conclusion

Applied Machine Learning in Healthcare is an active area of research. The increasingly widespread applicability of machine learning models necessitates the need for explanations to hold machine learning models accountable. While there is not much agreement on the meaning of interpretability in machine learning, there are a number of characteristics of interpretable models that researchers have discussed which can be used as a guide to create the requirements of interpretable models. The choice of interpretable models depends upon the application an use case for which explanations are required. Thus a critical application like prediction a patient's end of life may have much more stringent conditions for explanation fidelity as compared to just predicting costs for a procedure where getting the prediction right is much more important as compared to providing explanations. There are still a large number of questions that are unaddressed in the area of interpretable models and we envision that it will be an active area of research for the next few years.

Acknowledgment

The authors would like to thank Kiyana Zolfaghar, Vikas Kumar, Dr. Nicholas Mark, Rohan D'Souza and other members of the KenSci team for the fruitful and enlightening discussions and feedback on interpretability in machine learning.

References

[1] TR Addis. Towards an" expert" diagnostic system. *ICL Technical Journal*, 1:79–105, 1956.

[2] Muhammad Aurangzeb Ahmad, Zoheb Borbora, Jaideep Srivastava, and Noshir Contractor. Link prediction across multiple social networks. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 911–918. IEEE, 2010.

[3] Maruan Al-Shedivat, Avinava Dubey, and Eric P Xing. Contextual explanation networks. *arXiv preprint arXiv:1705.10301*, 2017.

[4] Jenna Burrell. How the machine thinks: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512, 2016.

[5] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730. ACM, 2015.

[6] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

[7] Pedro Domingos. The role of occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3(4):409–425, 1999.

[8] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.

[9] Alex A Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, 15(1):1–10, 2014.

[10] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[11] Puneet Gupta. Machine learning: The future of healthcare. *Harvard Sci. Rev.*, 2017.

[12] Sara Hajian, Francesco Bonchi, and Carlos Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2125–2126. ACM, 2016.

[13] Alvin D Jeffery, Laurie L Novak, Betsy Kennedy, Mary S Dietrich, and Lorraine C Mion. Participatory design of probability-based decision support tools for in-hospital nurses. *Journal of the American Medical Informatics Association*, 24(6):1102–1110, 2017.

[14] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):15, 2012.

[15] Samantha Krening, Brent Harrison, Karen M Feigh, Charles Lee Isbell, Mark Riedl, and Andrea Thomaz. Learning from explanations using sentiment and advice in rl. *IEEE Transactions on Cognitive and Developmental Systems*, 9(1):44–55, 2017.

[16] Todd Kulesza. Personalizing machine learning systems with explanatory debugging. PhD Dissertation, Oregon State University, 2014.

[17] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[18] Gunasekaran Manogaran and Daphne Lopez. A survey of big data architectures and machine learning algorithms in healthcare. *International Journal of Biomedical Engineering and Technology*, 25(2-4):182–211, 2017.

[19] Tim Miller. Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint arXiv:1706.07269*, 2017.

[20] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 2017.

[21] Mehran Mozaffari-Kermani, Susmita Sur-Kolay, Anand Raghunathan, and Niraj K Jha. Systematic poisoning attacks on and defenses for machine learning in healthcare. *IEEE Journal of Biomedical and Health Informatics*, 19(6):1893–1905, 2015.

[22] Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care. *Jama*, 309(13):1351–1352, 2013.

[23] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[24] Jesus Maria Pérez, Javier Muguerza, Olatz Arbelaitz, and Ibai Gurrutxaaga. A new algorithm to build consolidated trees: study of the error rate and steadiness. In *Intelligent Information Processing and Web Mining*, pages 79–88. Springer, 2004.

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[26] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014.

[27] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Detecting bias in black-box models using transparent model distillation. *arXiv preprint arXiv:1710.06169*, 2017.

[28] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. I read but don't agree: Privacy policy benchmarking using machine learning and the eu gdpr. In *Companion of The Web Conference 2018*, pages 163–166. International World Wide Web Conferences Steering Committee, 2018.

[29] Michael Tomasello. *The Cultural Origins of Human Cognition*. Harvard University Press, 2009.

[30] Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, 2016.

[31] Fulton Wang and Cynthia Rudin. Falling rule lists. In *Artificial Intelligence and Statistics*, pages 1013–1022, 2015.

[32] William Yang Wang. "Liar, Liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[33] Xin Yan and Xiaogang Su. *Linear Regression Analysis: Theory and Computing*. World Scientific, 2009.

Machines That Know Right And Cannot Do Wrong: The Theory and Practice of Machine Ethics

Louise A. Dennis and Marija Slavkovik

Abstract—**Machine ethics is an emerging discipline in Artificial Intelligence (AI) concerned with enabling autonomous intelligent systems to uphold the ethical, legal and societal norms of their environment. Why is machine ethics only developing as a field now, and what are its main goals and challenges? We tackle these questions and give a survey of state of the art implementations.**

"The fact that man knows right from wrong proves his intellectual superiority to the other creatures; but the fact that he can do wrong proves his moral inferiority to any creatures that cannot." – Mark Twain

I. Morality for machines

THE scenario is by now familiar: you are in a tunnel and your autonomous car has a break failure. There are workers on the road ahead. What should the car do? One option is to ram in the wall and possibly kill you, its owner and sole passenger. The other, to continue straight on its way and kill numerous road workers. Many questions are open regarding what the car should do, all subject of *machine ethics* [26].

The first challenge facing machine ethicists is what ethical conduct should an autonomous system exhibit and who gets to decide this. An equally important challenge is the one that we focus on here: How should an autonomous system be built and programmed so as to follow the ethical codex of choice? How can we do this in a way that allows a regulatory body to determine that the ethical behaviour described is the one exhibited? In summary, what does it mean to construct an artificial system that knows right from wrong and then ensure that it, unlike man in Mark Twain's quote, is unable to do wrong.

II. Why now?

AI has been an established research field since 1956 [31] but machine ethics, outside of science fiction, has emerged as a concern in the last decade. Why only now? At least two things have recently changed in how AI is used.

Powerful autonomous systems now share in our physical and e-space. Consider for example, industrial robots that have been in operation at least since the 80ies [27], and automated subway systems, which have been in operation for the past forty years[1]. Both of these types of machines have the capacity to seriously harm people and property, however they operate in a *work envelope*, a segregated space which only trained personel are allowed to enter. Machines that did share the space with people had no physical ability to do harm, such as automated pool cleaners. In contrast, machines like automated cars and assisted living devices have the ability do do harm and are not operating in a segregated environment.

Methods developed in AI have long been in use: *e.g.,* complex scheduling systems built using constraint satisfaction programming [33]. However, each of these AI systems have been domain and context specific. Any possible ethical, legal and societal issues that might arise from the use and deployment of the system could and had been handled during development. Today in contrast, particularly with machine learning applications, we see off-the shelf software and hardware available to any one to customize and deploy for an unpredictable variety of tasks in an unpredictable variety of contexts. Thus issues of machine "unethical behaviour" and impact can no longer be dealt with entirely in development.

III. Morality as a function of ability

Much has been said on whether an artificial agent, can be a moral agent, see *e.g.,* [17]. As with autonomy, we tend to refer to two different concepts: categorical morality for people, and degrees of morality for machines [35], [26].

Wallach and Allen [35, Chapter 2] distinguish between operational morality, functional morality, and full moral agency. An agent has operational morality when the moral significance of her actions are entirely scoped by the agent's designers. An agent has functional morality when the agent is able to make moral judgements when choosing an action, without direct human instructions.

Moor [26] distinguishes between agents with ethical impact, implicitly ethical, explicitly ethical and full moral agents. An agent has ethical impact if her operations increase or decrease the overall good in the world. A parallel can be drawn to [35]: implicitly ethical agents have operational morality, while explicitly ethical agents have functional morality. Dyrkolbotn et al. [15] further refine and formalise the concepts of implicitly and explicitly ethical agents by stipulating that implicitly ethical agents are those which do not use their autonomy to make moral judgements.

It is clearly better to build implicitly ethical artificial agents because their moral choices can be evaluated while the agent is built and assurances can be given about what the agent will do in a morally sensitive context. However, for agents whose context of operation is either unpredictable or too complex, explicit moral agency is the only design option [15].

L. Dennis was funded by EPSRC grants EP/L024845/1 (Verifiable Autonomy), EP/R026084/1 (Robotics and AI for Nuclear) and EP/R026092/1 (Future AI and Robotics Hub for Space).

[1] http://www.railjournal.com/index.php/metros/uitp-forecasts-2200km-of-automated-metros-by-2025.html

Having chosen what kind of artificial moral agent one needs, one has a choice between a bottom-up, top-down or a hybrid approach [36], [10]. In a top-down approach an existing moral theory is chosen and the agent is implemented with an ability to use this theory. In a bottom-up approach, the artificial agent is presented with examples of desirable and undesirable choices and she develops an algorithm by which to make moral judgements in unfamiliar circumstances. A hybrid approach uses elements of both the top-down and bottom-up. All of these approaches have advantages and disadvantages [10].

IV. Ethical Theories for Machines

Moral philosophy is concerned with developing moral theories, which should guide moral judgements. However the theories so far developed have a human locus, so not all can be trivially adapted for use by artificial agents. How can virtue ethics [22] for example, be used for an agent that can choose her reward function? Alternatively one might consider developing a new moral theory, specifically for machines. A (perhaps bad) example of such a theory are the Three Laws of Robotics of Asimov [4].

Ethical theories considered for use by artificial agents are: utilitarianism [21], Ross's ethical theory [30], and Kantianism [16]. Utilitarianism stipulates that all choices can be evaluated by the amount of good or bad (utilities) that they bring about. A moral agent needs to maximise the utility sum of her actions. 'W.D. Ross [30] argues that no absolute moral theory can be developed and suggests instead that a set of principles, or *prima facie* duties is used whenever possible: fidelity, reparation, gratitude, non-injury, harm-prevention, beneficence, self-improvement and justice.

Kant suggests that a moral agent follows a set of categorical imperatives which are maxims that are sufficiently virtuous to be used by everyone at every context. Here the principle of double effect should also be mentioned [25]. According to this principle (or doctrine), unethical actions can be sometimes permissible as a side effect of pursuing a moral action. Those same "bad" actions would not be permissible when they are the *means* to accomplishing the same moral action. In general, these theories are ones in which the intentions of the actor are important in determining the ethics of an action. A variation of these theories are ones in which actions themselves have ethical force. Deontic logics [20] that specify the actions an agent is obliged to take or prohibited from taking are well studied and supported by a variety of programming frameworks which have been applied to normative reasoning in general not just ethical reasoning.

V. Giving Machines the Capacity to Know Right

All machine reasoning systems can be viewed as ethical reasoning systems at some level of abstraction. We survey the key contribution systems that are explicitly ethical [26].

A. GenEth

The GenEth system [1] has two purposes. Firstly, it demonstrates how input from professional ethicists can be used, via a process of machine learning, to create a *principle of ethical action preference*. GenEth analyses a situation in order to determine its ethical features (*e.g.,* that physical harm may befall someone). These features then give rise to *prima facie* duties (to minimize or maximize that feature). In this theoretical framework GenEth is explicitly adopting Ross' theory of *prima facie* duties.

The principle of ethical action preference is used to compare two options: each option is assigned a score for each ethical feature, the scores are then used by the principle to determine the appropriate course of action based on which, duties are of more importance given the other duties effected. *E.g.,* the system might prefer an action which had worse consequences for privacy on the grounds it was better for safety.

GenEth can "explain" its decisions in terms of its preferences over duties – so it can state how two options compared on the various ethical features and refer to the statement of the principle. It is important to emphasize this feature of *explainability* particularly since GenEth uses machine learning as part of the process by which its ethical behaviour is determined. Machine learning systems, in general, are not particularly transparent to users, but some can be made so.

B. $\mathcal{DCEC}_{CL}$

Bringsjord et al. have a body of work [8], [9], developing the *deontic cognitive event calculus,* $\mathcal{DCEC}_{CL}$, in which various ethical theories can be expressed. A key motivation is a belief that ethical reasoning must necessarily be implemented at the operating system level. Concepts in the $\mathcal{DCEC}_{CL}$ are expressed in explicitly deontological terms – *i.e.,* as obligations, permissions and prohibitions.

An illustrative example of the $\mathcal{DCEC}_{CL}$ approach is the Akratic robot [9]. This considers a scenario in which a robot charged with guarding a prisoner of war must choose whether or not to retaliate with violence to an attack. [9] argues that the underlying robot architecture, into which the modules for self-defence and detainee management have been embedded, must be capable of ethical reasoning in order to predict and prevent ethical conflicts.

$\mathcal{DCEC}_{CL}$ uses automated reasoning to deduce ethical courses of action by reasoning explicitly about its obligations, prohibitions and so on. Automated reasoning, also referred to as automated theorem proving, has a long history in AI [29], with particular attention paid to implementations with high degrees of assurance. As a result automated reasoning with $\mathcal{DCEC}_{CL}$ can be considered *correct by virtue of the reasoning process* so long as the concepts supplied correctly capture the values of the community the system is designed to serve.

C. Ethical Governors

Arkin et. al [2], [3] outline the architecture for an *ethical governor* for automated targeting systems. This governor is charged with ensuring that any use of lethal force is governed by the "Law of War", the "Rules of Engagement". This initial work on was then re-implemented in a new setting of healthcare [32]. The governor is implemented as a separate module that intercepts signals from the underlying deliberative system and, where these signals involve lethality, engages in a process of *evidential reasoning* which amasses information about the situation in a logical form and then reasons using prohibitions and obligations. If any prohibitions are violated or obligations unfulfilled then the proposed action is vetoed.

The authors note that "it is a major assumption of this research that accurate target discrimination with associated uncertainty measures can be achieved despite the fog of war". It should be noted that throughout the literature on machine ethics there is an assumption seldom explicitly stated as it is in Arkin's work that complex, sometimes highly nuanced, information is available to the ethical reasoning system in order for it to make a determination. A key open area of research in machine ethics would seem to be the development of techniques for *ethical situation awareness*. The explicit use of evidential reasoning is an important step towards developing such techniques but only part of the story.

Unlike $\mathcal{DCEC}_{CL}$, the reasoning used by Arkin's ethical governors is not grounded in a formal logical theory. Ad-hoc reasoning techniques are therefore used rather than ones derived from automated theorem proving – as such deductions can not be assumed correct by virtue of the reasoning process.

D. Ethical Consequence Engines

Winfield et. al [34] have investigated systems based on the concept of an *Ethical Consequence Engine*. Ethical consequence engines are grounded in consequentialist theories of ethics, particularly utilitarianism. Like ethical governors, ethical consequence engines, pay attention to the ethical information upon which reasoning is based. Given they are using utilitarian ethics the question becomes one of generating appropriate utilities for each action.

The consequence engines use simulation to evaluate the impact of actions on the environment. In particular they simulate *not just* the actions of a robot itself but the activity of other agents in the environment. This allows the robot to determine not only if its actions have directly negative consequences (*e.g.,* colliding with a person) but if they have indirectly negative consequences *e.g.,* failing to intercept a person who might come into danger). The ethics implemented in each system thus has a distinctly Asimovian flavour, as directly acknowledged in [34]. The implemented ethical system can be seen as a combination of utilitarianism and Asimov's Laws.

E. ETHAN

The ETHAN system [13] was developed to investigate ethical decision making in exceptional circumstances. In ETHAN a *rational agent* [28] reasons about the ethical risks of plans proposed by an underlying planning system. The operation of reasoning in normal circumstances is assumed to be ethical by default (*i.e.,* that the agent is implicitly ethical), but in exceptional circumstances the system might need to make use of techniques such as planning or learning whose behaviour is difficult to analyse in advance.

[13] considers the case of a planning system that returns candidate plans to the agent which are annotated with context specific ethical concerns. These concerns are then reasoned about using a priority-based context specific *ethical policy* that prefers plans violating lower priority concerns to plans violating higher priority concerns and, where two plans violate concerns of the same priority, prefers the plan violating the fewest concerns. As with GENETH, ETHAN's ethics are based on Ross's *prima facie* duties [30] and ETHAN's ethical principles can be considered broadly similar to GENETH's ethical features.

F. HERA

The hybrid ethical reasoning agent (HERA) system [24] uses a model theoretic approach to investigate the implementation of different ethical theories. Its primary focus has been constructing a rich framework that can express both Utilitarian and Kantian/Deontological systems – in particular the categorical imperative [6] and the principle of double effect [5].

For each action available to it, HERA builds a model depicting the overall utility of the action, as well as whose utilities are affected (positively or negatively) and which agents are *ends* of the action and which are affected as *means* to those ends. These models have a formal basis allowing automated reasoning to determine whether some logical formula is satisfied by the model, so again this reasoning can be considered correct by virtue of the reasoning process.

In the case of utilitarianism HERA compares all models and selects the one with the highest overall utility. In the case of the categorical imperative and principle of double-effect it constructs a logical formula expressing the ethical constraints and then vetoes models which do not satisfy the formula.

VI. ENSURING A MACHINE CAN NOT DO WRONG

Formal verification is the process of assessing whether a formal specification is satisfied on a particular formal description of a system. For a specific logical property, φ, there are many different approaches to this [18], [12], [7], ranging from deductive verification against a logical description of the system ψ_S (*i.e.,*$\vdash \psi_S \rightarrow \varphi$) to the algorithmic verification of the property against a model of the system, M (*i.e.,*$M \models \varphi$). The latter has been extremely successful in Computer Science and AI, primarily through the *model checking* approach [11]. This takes a model of the system in question, defining all the model's possible executions, and then checks a logical property against this model.

The approach most often applied to the verification of machine ethics is a model-checking approach for the verification of agent-based autonomous systems outlined in [19] which considers the decision taken by the system given any combination of incoming information. This methodology adapts well if we can implement an ethical decision agent on top of an underlying autonomous system which accepts processed ethical information as input. We note that this is the architecture adopted in most of the systems we have described. A model-checker can then verify that such a system always chooses options that align with a given code of ethics based on the information that it has. This approach has been applied both to the verification of ETHAN programs [13] and to the verification of ethical consequence engines [14].

In ETHAN programs the emergency planning system was replaced by a random component that generated plans annotated as violating some combination of ethical concerns. The model-checking process then ensured that all such combinations were considered. Given a ranking of concerns according to some ethical policy the verification was able to show that a plan was only selected by the system if all other plans were annotated as violating some more serious ethical concern. In [14] a simplified model of the ethical consequence engine was constructed on a 5x5 grid. This was used to check the decision making as in the ETHAN system. In an extension, a probabilistic model of the human behaviour was also created in order to use a probabilistic model-checker (PRISM [23]) to generate probabilities that the robot would successfully "rescue" a human given any combination of "human" movement on the grid. The results of this verification differed greatly from the probabilities generated through experimental work in a large part because the model used in verification differed significantly, in terms of the environment in which the robot operated to the environment used experimentally.

HERA and $\mathcal{DCEC}_{CL}$ use formal logical reasoning in order to make ethical choices – model checking in HERA and theorem proving in $\mathcal{DCEC}_{CL}$. For simple models/formulae it is easy to rely on the correctness of this reasoning to yield correct results but we note that for more complex models this is more challenging. Even in systems that perform ethical reasoning that is correct by virtue of the reasoning process, it may be necessary to verify some "sanity" properties.

VII. CONCLUSIONS

We here attempted to survey the current state of the art in the implementation and verification of machine ethics having noted that, unlike human reasoning, we require machine ethical reasoners not only to know which is the correct action, but also then act in accordance with that knowledge. We have restricted ourselves to explicitly ethical systems which reason about ethical concepts as part of the system operation. While the field of practical machine ethics is still in its infancy, it is thus possible to see some clear convergence in approaches to implementation and consensus about the need for strong assurances of correct reasoning.

REFERENCES

[1] M. Anderson and S. Leigh Anderson. Geneth: A general ethical dilemma analyzer. In *Proceedings of the 28th AAAI Conference on AI, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 253–261, 2014.

[2] R.C. Arkin, P. Ulam, and B. Duncan. An Ethical Governor for Constraining Lethal Action in an Autonomous System. Technical report, Mobile Robot Laboratory, College of Computing, Georgia Tech., 2009.

[3] R.C. Arkin, P. Ulam, and A. R. Wagner. Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. *Proc. of the IEEE*, 100(3):571–589, 2012.

[4] I. Asimov. *I, Robot*. Gnome Press, 1950.

[5] M.M. Bentzen. *The principle of double effect applied to ethical dilemmas of social robots*, pages 268–279. IOS Press, 2016.

[6] M.M. Bentzen and F. Lindner. A formalization of kant's second formulation of the categorical imperative. *CoRR*, abs/1801.03160, 2018.

[7] R. S. Boyer and J. Strother Moore, editors. *The Correctness Problem in Computer Science*. Academic Press, London, 1981.

[8] S. Bringsjord, K. Arkoudas, and P. Bello. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21(4):38–44, 2008.

[9] S. Bringsjord, N Sundar, D. Thero, and M. Si. Akratic robots and the computational logic thereof. In *Proc. of the IEEE 2014 Int. Symposium on Ethics in Engineering, Science, and Technology*, pages 7:1–7:8, Piscataway, NJ, USA, 2014.

[10] V. Charisi, L.A. Dennis, M. Fisher, R. Lieck, A. Matthias, M. Slavkovik, J. Sombetzki, A.F.T. Winfield, and R. Yampolskiy. Towards moral autonomous systems. *CoRR*, abs/1703.04741, 2017.

[11] E. Clarke, O. Grumberg, and D. Peled. *Model Checking*. MIT Press, 1999.

[12] R. A. DeMillo, R. J. Lipton, and A.J. Perlis. Social Processes and Proofs of Theorems of Programs. *ACM Communications*, 22(5):271–280, 1979.

[13] L. A. Dennis, M. Fisher, M. Slavkovik, and M. P. Webster. Formal Verification of Ethical Choices in Autonomous Systems. *Robotics and Autonomous Systems*, 77:1–14, 2016.

[14] L. A. Dennis, M. Fisher, and A. F. T. Winfield. Towards Verifiably Ethical Robot Behaviour. In *Proceedings of AAAI Workshop on AI and Ethics*, 2015.

[15] S. Dyrkolbotn, T. Pedersen, and M. Slavkovik. On the distinction between implicit and explicit ethical agency. In *AAAI/ACM Conference on AI, Ethics and Society*, New Orleans, USA, 2018.

[16] J. W. Ellington. *Translation of: Grounding for the Metaphysics of Morals: with On a Supposed Right to Lie because of Philanthropic Concerns by Kant, I. [1785]*. Hackett Publishing Company, 1993.

[17] A. Etzioni and O. Etzioni. Incorporating ethics into artificial intelligence. *The Journal of Ethics*, pages 1–16, 2017.

[18] J. H. Fetzer. Program Verification: The Very Idea. *ACM Communications*, 31(9):1048–1063, 1988.

[19] M. Fisher, L. Dennis, and M. Webster. Verifying Autonomous Systems. *ACM Communications*, 56(9):84–93, 2013.

[20] D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre, editors. *Handbook of Deontic Logic and Normative Systems*. College Publications, London, UK, 2013.

[21] J.C. Harsanyi. Rule utilitarianism and decision theory. *Erkenntnis (1975-)*, 11(1):25–53, 1977.

[22] R. Hursthouse and G. Pettigrove. Virtue ethics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

[23] M. Kwiatkowska, G. Norman, and D. Parker. PRISM: Probabilistic Symbolic Model Checker. In *Proc. 12th Int. Conf. Modelling Techniques and Tools for Computer Performance Evaluation (TOOLS)*, volume 2324 of *LNCS*, 2002.

[24] F. Lindner and M.M. Bentzen. The hybrid ethical reasoning agent IMMANUEL. In *Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2017, Vienna, Austria, March 6-9, 2017*, pages 187–188, 2017.

[25] A. McIntyre. Doctrine of double effect. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter edition, 2014.

[26] J. H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21, 2006.

[27] S.Y. Nof. *Handbook of Industrial Robotics*. Number v. 1 in Electrical and electronic engineering. Wiley, 1999.

[28] A. S. Rao and M. P. Georgeff. BDI Agents: From Theory to Practice. *Proc. of the First International Conference on Multiagent Systems*, 95:312–319, 1995.

[29] A. Robinson and A. Voronkov, editors. *Handbook of Automated Reasoning*. Elsevier Science Publishers B. V., 2001.

[30] W.D. Ross. *The Right and the Good*. Oxford University Press, 1930.

[31] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2009.

[32] J. Shim and R. C. Arkin. An Intervening Ethical Governor for a Robot Mediator in Patient-Caregiver Relationships. In M. I. Aldinhas Ferreira et al., editor, *A World with Robots: International Conference on Robot Ethics: ICRE 2015*, pages 77–91. Springer Int. Publishing, 2017.

[33] H. Simonis. Constraints in computational logics. chapter Building Industrial Applications with Constraint Programming, pages 271–309. Springer-Verlag New York, Inc., 2001.

[34] D. Vanderelst and A. Winfield. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, 2017.

[35] W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2008.

[36] W. Wallach, C. Allen, and I. Smit. Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI Society*, 22(4):565–582, 2008.

Diffusion Mechanism Design in Social Networks

Dengji Zhao

***Abstract*—In this article, we introduce the diffusion mechanisms that we have proposed [1], [2]. We consider a market where a seller sells multiple units of a commodity in a social network. Each node/buyer in the social network can only directly communicate with her neighbours, i.e. the seller can only sell the commodity to her neighbours if she could not find a way to inform other buyers. We have designed a novel promotion mechanism that incentivizes all buyers, who are aware of the sale, to invite all their neighbours to join the sale, even though there is no guarantee that their efforts will be paid. While traditional sale promotions such as sponsored search auctions cannot guarantee a positive return for the advertiser (the seller), our mechanism guarantees that the seller's revenue is better than not using our promotion mechanism. More importantly, the seller does not need to pay if the promotion is not beneficial to her. In this article, we briefly introduce our mechanism in a simple setting and highlight some open problems for further investigations.**

***Index Terms*—Mechanism design, information diffusion, revenue maximisation, algorithmic game theory**

I. Introduction

MARKETING is one of the key operations for a service or product to survive. To do that, companies often use newspapers, tv, social media, search engines to do advertisements. Indeed, most of the revenue of social media and search engines comes from paid advertisements. According to Statista, Google's ad revenue amounted to almost 79.4 billion US dollars in 2016. However, whether all the advertisers actually benefit from their advertisements is not clear and is difficult to monitor. Although most search engines use market mechanims like generalised second price auctions to allocate advertisements and only charge the advertisers when users click their ads, not all clicks lead to a purchase [3], [4]. That said, the advertisers may pay user clicks that have no value to them.

In order to guarantee that a seller never loses from using advertising, we have proposed novel advertising mechanisms without using third-party advertising platforms for the seller (to sell services or products) that do not charge the seller unless the advertising brings revenue-increase for the seller [1], [2], [5]. We model all potential buyers of a service/product as a large social network where each buyer is linked with some other buyers (known as neighbours). The seller is also located somewhere in the social network. Before the seller finds a way to inform more buyers about her sale, she can only sell her products to her neighbours. In order to attract more buyers to increase her revenue, the seller may pay to advertise the sale via newspapers, social media, search engines etc. to reach/inform more potential buyers in the social network. However, if the advertisements do not bring any valuable buyers, the seller loses the investment on the advertisements.

Our advertising mechanism does not rely on any third party such as newspapers or search engines to do the advertisements. The mechanism is owned by the seller. The seller just needs to invite all her neighbours to join the sale, then her neighbours will further invite their neighbours and so on. In the end, all buyers in the social network will be invited to participate in the sale. Moreover, all buyers are not paid in advance for their invitations and they may not get paid if their invitations are not beneficial to the seller. Although some buyers may never get paid for their efforts in the advertising, they are still incentivized to do so, which is one of the key features of our advertising mechanism. This significantly differs from existing advertising mechanisms used on the Internet.

More importantly, our advertising mechanism not only incentivizes all buyers to do the advertising, but also guarantees that the seller's revenue increases. That is, her revenue is never worse than the revenue she can get if she only sells the items to her neighbours.

Maximising the seller's revenue has been well studied in the literature, but the existing models assumed that the buyers are all known to the seller and the aim is to maximize the revenue among the fixed number of buyers. Given the number of buyers is fixed, if we have some prior information about their valuations, Myerson [6] proposed a mechanism by adding a reserve price to the original Vickrey-Clarke-Groves (VCG) mechanism. Myerson's mechanism maximises the seller's revenue, but requires the distributions of buyers' valuations to compute the reserve price. Without any prior information about the buyers' valuations, we cannot design a mechanism that can maximise the revenue in all settings (see Chapter 13 of [7] for a detailed survey). Goldberg et al. [8], [9] have considered how to optimize the revenue for selling multiple homogeneous items such as digital goods like software (unlimited supply). Especially, the seller can choose to sell less with a higher price to gain more.

In terms of incentivizing people to share information (like buyers inviting their neighbours), there also exists a growing body of work [10], [11], [12], [13]. Their settings are essentially different from ours however. They considered either how information is propagated in a social network or how to design reward mechanisms to incentivize people to invite more people to accomplish a challenge together. The mechanism designed by the MIT team under the DARPA Network Challenge (2009) is a nice example, where they designed a novel reward mechanism to share the award if they win the challenge. Thier mechanism attracted many people via social network to join the team, which eventually helped them to win the challenge [12].

Dr. Dengji Zhao is a tenure-track Assistant Professor at ShanghaiTech University, China. (e-mail: dengji.zhao@gmail.com)

II. The Model

We consider a seller s sells $\mathcal{K} \geq 1$ items in a social network. In addition to the seller, the social network consists of n nodes denoted by $N = \{1, \cdots, n\}$, and each node $i \in N \cup \{s\}$ has a set of neighbours denoted by $r_i \subseteq N \cup \{s\}$. Each $i \in N$ is a buyer of the $\mathcal{K}$ items.

For simplicity, we assume that the $\mathcal{K}$ items are homogeneous and each buyer $i \in N$ requires at most one unit of the item and has a valuation $v_i \geq 0$ for one or more units.

Without any advertising, seller s can only sell to her neighbours r_s as she is not aware of the rest of the network and the other buyers also do not know the seller s. In order to maximize s's profit, it would be better if all buyers in the network could join the sale.

Traditionally, the seller may pay some of her neighbours to advertise the sale to their neighbours, but the neighbours may not bring any valuable buyers and cost the seller money for the advertisement. Therefore, our goal here is to design a kind of cost-free advertising mechanism such that all buyers who are aware of the sale are incentivized to invite all their neighbours to join the sale with no guarantee that their efforts will be paid.

Let us first formally describe the model. Let $\theta_i = (v_i, r_i)$ be the *type* of buyer $i \in N$, $\theta = (\theta_1, \cdots, \theta_n)$ be the type profile of all buyers and θ_{-i} be the type profile of all buyers except i. θ can also be represented by (θ_i, θ_{-i}). Let Θ_i be the type space of buyer i and Θ be the type profile space of all buyers.

The advertising mechanism consists of an *allocation policy* π and a *payment policy* x. The mechanism requires each buyer who is aware of the sale to report her valuation to the mechanism and invite all her neighbours to join the sale. Let v'_i be the valuation report of buyer i and $r'_i \subseteq r_i$ be the neighbours i has invited. Let $\theta'_i = (v'_i, r'_i)$ and $\theta' = (\theta'_1, \cdots, \theta'_n)$, where $\theta'_j = nil$ if j has never been invited by any of her neighbours r_j or j does not want to participate. Given the action profile θ' of all buyers, $\pi_i(\theta') \in \{0, 1\}$, 1 means that i receives one item, while 0 means i does not receive any item. $x_i(\theta') \in \mathbb{R}$ is the payment that i pays to the mechanism, $x_i(\theta') < 0$ means that i receives $|x_i(\theta')|$ from the mechanism.

Different from the traditional mechanism design settings, in this model, we want to incentivize buyers to not only just report their valuations truthfully, but also invite all their neighbours to join the sale/auction (the advertising part). Therefore, we extend the definition of incentive compatibility to cover the invitation of their neighbours. Specifically, a mechanism is incentive compatible (or truthful) if for all buyers who are invited by at least one of their neighbours, reporting their valuations truthfully to the mechanism and further inviting all their neighbours to join the sale is a dominant strategy.

III. The Diffusion Mechanism

In this section, we review the diffusion mechanism proposed by Zhao et al. [2] for the case of $\mathcal{K} = 1$. The essence of our mechanism is that a buyer is rewarded for advertising the sale only if her invitations increase social welfare, and the reward guarantees that inviting all neighbours is a dominant strategy for all buyers.

The diffusion mechanism is outlined below:

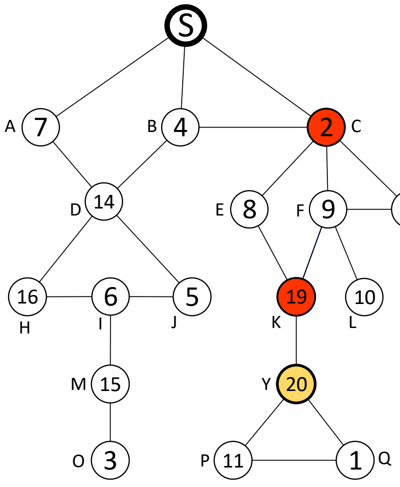

Fig. 1. A running example of the information diffusion mechanism, where the seller s is located at the top of the graph and is selling one item, the value in each node is the node's private valuation for receiving the item, and the lines between nodes represent neighbourhood relationship. Node Y is the node with the highest valuation and C, K are Y's diffusion critical buyers.

Information Diffusion Mechanism (IDM)

1) Given a feasible action profile θ', identify the buyer with the highest valuation, denoted by i^*.
2) Find all *diffusion critical buyers* of i^*, denoted by C_{i^*}. $j \in C_{i^*}$ if and only if without j's action θ'_j, there is no invitation chain from the seller s to i^* following θ'_{-j}, i.e. i^* is not able to join the sale without j.
3) For any two buyers $i, j \in C_{i^*} \cup \{i^*\}$, define an order $\succ_{i^*}$ such that $i \succ_{i^*} j$ if and only if all invitation chains from s to j contain i.
4) For each $i \in C_{i^*} \cup \{i^*\}$, if i receives the item, the payment of i is the highest valuation report without i's participation. Formally, let N_{-i} be the set of buyers each of whom has an invitation chain from s following θ'_{-i}, i's payment to receive the item is $p_i = max_{j \in N_{-i} \wedge \theta'_j \neq nil} v'_j$.
5) The seller initially gives the item to the buyer i ranked first in $C_{i^*} \cup \{i^*\}$, let $l = 1$ and repeat the following until the item is allocated.
   - if i is the last ranked buyer in $C_{i^*} \cup \{i^*\}$, then i receives the item and her payment is $x_i(\theta') = p_i$;
   - else if $v'_i = p_j$, where j is the $(l+1)$-th ranked buyer in $C_{i^*} \cup \{i^*\}$, then i receives the item and her payment is $x_i(\theta') = p_i$;
   - otherwise, i passes the item to buyer j and i's payment is $x_i(\theta') = p_i - p_j$, where j is the $(l+1)$-th ranked buyer in $C_{i^*} \cup \{i^*\}$. Set $i = j$ and $l = l + 1$.
6) The payments of all the rest buyers are zero.

Figure 1 shows a social network example. Without any advertising, the seller can only sell the item among nodes A, B and C, and her revenue cannot be more than 7. If A, B and C invite their neighbours, these neighbours further invite their neighbours and so on, then all nodes in the social network will be able to join the sale and the seller may receive a revenue as high as the highest valuation of the social network which is 20.

Let us run IDM on the social network given in Figure 1. Assume that all buyers report their valuations truthfully and invite all their neighbours, IDM runs as follows:

- Step (1) identifies that the buyer with the highest valuation is Y, i.e. $i^* = Y$.
- Step (2) computes $C_{i^*} = \{C, K\}$.
- Step (3) gives the order of $C_{i^*} \cup \{i^*\}$ as $C \succ_{i^*} K \succ_{i^*} i^*$.
- Step (4) defines the payments p_i for all nodes in $C_{i^*} \cup \{i^*\}$, which are $p_C = 16$, $p_K = 17$ and $p_Y = 19$, the highest valuation without C, K and Y's participation respectively.
- Step (5) first gives the item to node C; C is not the last ranked buyer in $C_{i^*} \cup \{i^*\}$ and $v_C \neq p_K$, so C passes the item to K and her payment is $p_C - p_K = -1$; K is not the last ranked buyer, but $v_K = p_Y$, therefore K receives the item and pays p_K.
- All the rest of the buyers, including Y, pay nothing.

In the above example, IDM allocates the item to node K and K pays 17, but s does not receive all the payment, and she pays C an amount of 1 for the advertising. Therefore, the seller receives a revenue of 16 from IDM, which is more than two times the revenue she can get without any advertising. Note that only buyer C is rewarded for the information propagation as the other buyers are not critical for inviting K.

A. Properties of the Diffusion Mechanism

Firstly, we can show that for all buyers who are invited by at least one of their neighbours, reporting their valuations truthfully to the mechanism (i.e. the seller) and further inviting all their neighbours to join the sale is a dominant strategy. Secondly, all buyers' utilities are non-negative, i.e. they are not forced to join the sale. Lastly, the seller's revenue is greater than or equal to the revenue she could get under the second price auction (Vickrey auction) among her neighbours only. All the properties together solve the dilemma that the seller has faced with the traditional advertising platforms such as search engines.

IV. Open Problems

Mechanism design in social networks is a very promising research direction, which has not been studied before in the literature of game theory. It also has a broader class of applications around the digital economy and the sharing economy. There are many open problems worth further investigations:

- Diffusion mechanisms for combinatorial settings: we have only looked at simple valuation settings. Whether our methods can be easily extended to more complex settings is an open question. As we have seen from [2], it is already very challenging to move from selling single-item setting to selling multiple-item setting.
- When diffusion is costly: we have assumed that information propagation is not costly, but in real-world applications, users might hesitate to do so, as propagating sale information to their friends might ruin their friendship. If diffusion is costly, can we still guarantee that the seller's revenue is non-decreasing with a diffusion mechanism? We have also considered transfer cost of the items in the network, which is not diffusion cost [5].
- In our setting, we also assumed that the seller is the market owner and she has the whole network structure (after the propagation). Since the seller is aware of the whole network, she can ignore paying other buyers and directly does transactions with the highest buyers. Moreover, buyers may not be confident to reveal their friendship to the seller, which is an important privacy concern in practice.
- Last but not least, buyers can create dummy friends to increase their payments, which is already a very hard problem in classical mechanism design settings [14]. Solving the challenge in our settings seems even harder.

Acknowledgment

The author would like to thank Bin Li, Junping Xu, Dong Hao, Tao Zhou and Nicholas R. Jennings for their great contribution on the work presented in this article.

References

[1] B. Li, D. Hao, D. Zhao, and T. Zhou, "Mechanism design in social networks," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 586–592.

[2] D. Zhao, B. Li, J. Xu, D. Hao, and N. R. Jennings, "Selling multiple items via social networks," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, 2018, pp. 68–76.

[3] B. Edelman, M. Ostrovsky, and M. Schwarz, "Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords," *American Economic Review*, vol. 97, no. 1, pp. 242–259, 2007.

[4] H. R. Varian, "Online ad auctions," *The American Economic Review*, vol. 99, no. 2, pp. 430–434, 2009.

[5] B. Li, D. Hao, D. Zhao, and T. Zhou, "Customer sharing in economic networks with costs," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, 2018, pp. 368–374.

[6] R. B. Myerson, "Optimal auction design," *Mathematics of Operations Research*, vol. 6, no. 1, pp. 58–73, 1981.

[7] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic game theory*. Cambridge University Press Cambridge, 2007, vol. 1.

[8] A. V. Goldberg, J. D. Hartline, and A. Wright, "Competitive auctions and digital goods," in *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2001, pp. 735–744.

[9] A. V. Goldberg and J. D. Hartline, "Competitive auctions for multiple digital goods," in *European Symposium on Algorithms*. Springer, 2001, pp. 416–427.

[10] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[11] E. M. Rogers, *Diffusion of innovations*. Simon and Schuster, 2010.

[12] G. Pickard, W. Pan, I. Rahwan, M. Cebrian, R. Crane, A. Madan, and A. Pentland, "Time-critical social mobilization," *Science*, vol. 334, no. 6055, pp. 509–512, 2011.

[13] Y. Emek, R. Karidi, M. Tennenholtz, and A. Zohar, "Mechanisms for multi-level marketing," in *Proceedings of the 12th ACM conference on Electronic commerce*. ACM, 2011, pp. 209–218.

[14] M. Yokoo, "False-name bids in combinatorial auctions," *SIGecom Exchanges*, vol. 7, no. 1, pp. 48–51, 2007.

Selected Ph.D. Thesis Abstracts

This Ph.D. thesis abstracts section selected theses defended during the period of August 2017 to July 2018. These submissions cover numerous research and applications under intelligent informatics such as robotics, fuzzy systems, semantic query processing system, recommender system, bioinformatics, image processing, transfer learning, machine learning in healthcare, cyber-physical systems, game theory and text mining, etc. Fig.1 presents the word cloud of those abstracts.

Fig. 1. Word Cloud of the Ph.D Thesis Abstracts

Using Rules of Thumb to Repair Inconsistent Knowledge

Elie Merhej
elie.merhej@ugent.be
Ghent University, Belgium

AN important challenge that arises when trying to create a knowledge base from raw data is the inconsistency introduced by the integrity constraints that are imposed. The goal is then to restore consistency in the knowledge base. A common approach to achieve this goal is to find some sort of minimal repair. While this strategy is reasonable in the absence of any background knowledge, in real-world applications, additional knowledge about the system being modelled is often accessible. In this thesis, we study the use of such additional knowledge to repair inconsistencies in different types of systems. We encode this knowledge in the form of rules of thumb that act as ”soft constraints” in a system.

First, we study the impact of using rules of thumb to repair inconsistencies that are found in taxonomies that were automatically extracted from text corpora found on the web. We use Markov logic to encode this problem and propose MAP inference as a base method to generate minimal repairs. We encode dependencies between taxonomy facts in the form of rules of thumb, such as: ”if a given fact is wrong then all facts that have been extracted from the same sentence are also likely to be wrong”. We show that, by adding these rules, we generate more accurate repairs than minimal repairs.

Second, we introduce a rules of thumb approach to repair inconsistent answer set programs. Answer set programming (ASP) is a form of declarative programming that is used to model various systems. We consider the scenario where additional knowledge that could be encoded as rules of thumb is available about the studied domain, but no training data is available to learn how these rules interact. The main problem we address is whether we can still aggregate the rules of thumb in the absence of training data in order to generate more plausible repairs than minimal repairs. In addition to standard aggregation techniques, we present a novel statistical approach that assigns weights to these rules of thumb, by sampling from a pool of possible repairs. We show in our experiments that our Z-score approach outperforms all the other repair methods in terms of F_1 score and Jaccard index, including the minimal repair approach.

Third, we tackle the problem of repairing inconsistencies that arise when multiple treatments are simultaneously needed for comorbid patients. In this application, we encode preferences as rules of thumb that contain information in the form of drug-drug interactions. We show in a case study encoded in ASP that this method generates more preferred treatments than standard approaches. The second method we propose to find treatments for patients with comorbid diseases is a fully data driven approach using word-based and phrase-based alignment methods. In this approach, we explore the use of rules of thumb that incorporates drug interactions penalty and procedure popularity scores. We show that the combined treatments that are found when adding rules of thumb to these word-based and phrase-based alignment methods are more plausible than using standard translation approaches.

Optimization of Semantic Cache Query Processing System

Munir Ahmad
munirahmad83@gmail.com
Center for Distributed and Semantic Computing, Capital University of Science and Technology, Islamabad. Pakistan

HIGH availability and low latencies of data are major requirements in accessing contemporary large and networked databases. However, it becomes difficult to achieve high availability and reduced data access latency with unreliable connectivity and limited bandwidth. These two requirements become crucial in ubiquitous environment when data is required all the times and everywhere. Cache is one of the promising solutions to improve availability and reduce latencies of remotely accessed data by storing and reusing the results of already processed similar queries. Conventional cache lacks in partial reuse of already accessed data, while semantic cache overcomes the limitation of conventional cache by reusing the data for partial overlapped queries by storing description of queries with results. There is a need of an efficient cache system to improve the availability, reduce the data access latencies and the network traffic by reusing the already stored results for fully and partially overlapped queries. An efficient cache system demands efficient query processing and cache management. In this study, a qualitative benchmark with four qualities as Accuracy, Increased Data Availability, Reduced Network Traffic and Reduced Data Access Latency is proposed to evaluate a semantic cache system, especially from query processing point of view. The qualitative benchmark is then converted into six quantitative parameters (Semantics and Indexing Structure IS, Generation of Amending Query GoAQ, Zero Level Rejection ZLR, Predicate Matching, SELECT_CLAUSE Handling, Complexity of Query Matching CoQM) that help in measuring the efficiency of a query processing algorithm. As the result of evaluation, it is discovered that existing algorithms for query trimming can be optimized. Architecture of a semantic cache system is proposed to meet the benchmark criteria. One of the important deficiencies observed in the existing system is the storage of query semantics in segments (indexing of the semantics) and the organization of these segments. Therefore, an appropriate indexing scheme to store the semantics of queries is needed to reduce query matching time. In the existing indexing schemes the number of segments grows faster than exponential, i.e., more than 2^n. The semantic matching of a user query with number of segments more than 2^n will be exponential and not feasible for a large value of n. The proposed schema-based indexing scheme is of polynomial time complexity for the matching process. Another important deficiency observed is the large complexity of query trimming algorithm which is responsible to filter the semantics of incoming query into local cache and remote query. A rule based algorithm is proposed for query trimming that is faster and less complex than existing satisfiability/implication algorithms. The proposed trimming algorithm is more powerful in finding the hidden/implicit semantics, too. The significance of the proposed algorithms is justified by case studies in comparison with the previous algorithms and correctness is tested by implementing a prototype. The final outcomes revealed that the proposed scheme has achieved sufficient accuracy, increased availability, reduced network traffic, and reduced data access latency.

INVESTIGATING PROTEIN SEMANTIC SIMILARITY MEASUREMENT AND ITS CORRELATION WITH SEQUENCE SIMILARITY

Najmul Ikram
najmalikram@yahoo.com
Capital University of Science and Technology Islamabad, Pakistan

PROTEIN sequence similarity is commonly used to compare proteins and to search for proteins similar to a query protein. With the growing use of biomedical ontologies, especially Gene Ontology (GO), semantic similarity between ontology terms, proteins and genes is getting attention of researchers. Protein semantic similarity measurement has many applications in bioinformatics, including prediction of protein function and protein-protein interactions. Semantic similarity measures were initially proposed by Resnik, Jiang and Conrath, and Lin. Recent measures include Wang and AIC. The question whether the semantic similarity has strong correlation with sequence similarity, has been addressed by some authors. It has been reported that such correlation exists, and it has been used for the evaluation of semantic similarity computation methods as well as for protein function prediction. We investigate the correlation between semantic similarity and sequence similarity through graphs, Persons correlation coefficient and example proteins. We find that there is no strong correlation between the two similarity measures. Pearsons correlation coefficient is not sufficient to explain the nature of this relationship, if not accompanied by graph analysis. We find that there are several pairs with low sequence similarity and high semantic similarity, but very few pairs with high sequence similarity and low semantic similarity. Interestingly, the correlation coefficient depends only on the number of common GO terms in proteins under comparison. We propose a novel method SemSim for semantic similarity measurement. It addresses the limitations of existing methods, and computes similarity in two steps. In the first step, SimGIC like approach is used where contribution of common ancestors is divided by contribution of all ancestors. In the second step, we use two new factors: Specificity computed from ontology based information content, and Uniqueness computed from annotation based information content. The final result, after applying these two factors, makes clear distinction between the generalized and specialized terms. When semantic similarity is used for searching proteins from large databases, the speed issue becomes significant. To search for proteins similar to a query protein having m annotations, from the database of p proteins, p X m X n X g comparisons would be required. Here n is the average annotations per protein, g is the complexity of GO term similarity computation algorithm, and it is assumed that each term of one protein is compared with each term of the other. We propose a method SimExact that is suitable for high speed searching of semantically similar proteins. Although SimExact works on common terms only, our experiments show that it gives correct results required for protein semantic searching. SimExact can be used as a pre processor, generating candidate list for the existing methods, which proceed for further computation. We provide online tool that generates a ranked list of the proteins similar to a query protein, with a response time of less than 8 seconds in our setup. We use SimExact to search for protein pairs having high disparity between semantic similarity and sequence similarity. SimExact makes such searches possible, which would be NP-hard otherwise.

TIME-EFFICIENT VARIANTS OF TWIN SUPPORT VECTOR MACHINE WITH APPLICATIONS IN IMAGE PROCESSING

Pooja Saigal

saigal.pooja.sau@gmail.com
South Asian University, India

HUMAN beings can display intelligent behavior by learning from their experiences. The aim of learning is to generalize well, which essentially means to establish similarity between situations, so that the rules which are applicable in one situation can be applied or extended to other situations. Machine learning enables a machine to learn from empirical data and builds models to make reliable future predictions. It is categorized as supervised and unsupervised learning. Support Vector Machines and Twin Support Vector Machine (TWSVM) are distinguished works in supervised learning. This research work attempts to develop machine learning algorithms which could deliver better results than well-established methodologies. Our focus is on development of time-efficient learning algorithms, with good generalization ability, and to apply them for image processing tasks.

To improve the time complexity of nonparallel-hyperplane classifiers, this thesis first proposes a set of algorithms termed as Improvements on ν-Twin Support Vector Machine. The first version of our classification algorithm solves an efficient, smaller-sized quadratic programming problem (QPP) and an unconstrained minimization problem (UMP), instead of solving a pair of expensive QPPs. Second (and faster) version modifies first problem as minimization of unimodal function, for which line search methods can be used. Experimental results proved that proposed algorithms have good generalization ability and are extended to handle multi-category classification problems. Two more classifiers i.e. Angle-based Twin Parametric-Margin Support Vector Machine (ATP-SVM) and Angle-based Twin Support Vector Machine (ATWSVM), have been proposed, which aim to maximize the angle between normal vectors to the two nonparallel-hyperplanes, so as to generate larger separation between the two classes. ATP-SVM solves only one modified QPP with fewer representative patterns and avoids explicit computation of matrix inverse in the dual problem. This improves learning time of our algorithm. ATWSVM is a generic algorithm to improve efficiency of any existing binary nonparallel-hyperplane classifier.

This thesis proposes Ternary Support Vector Machine to separate data belonging to three classes and its multi-category classification algorithm, Reduced Tree for Ternary Support Vector Machine. Here, classes are organized in the form of ternary tree. Most of the real world problems deal with multiple classes, so this work proposes Ternary Decision Structure and Binary Tree of classifiers, that can extend existing binary classifiers to multi-category framework. They are more efficient than the classical multi-category classification approaches. This work proposes development of unsupervised clustering algorithm termed as Tree-based Localized Fuzzy Twin Support Vector Clustering (Tree-TWSVC), which recursively builds a cluster model as a Binary Tree. Here, each node comprises of a novel classifier termed as Localized Fuzzy TWSVM. Tree-TWSVC has efficient learning time, achieved due to tree structure and its formulation leads to solving a series of system of linear equations.

Extensive experiments have been carried out to prove the efficacy of proposed algorithms using synthetic and benchmark real-world datasets. Our algorithms have outperformed state-of-the-art methods and results presented in the thesis demonstrate their effectiveness and applicability. Our algorithms have been applied to perform image processing tasks like content based image retrieval, image segmentation, handwritten digit recognition. (http://sau.int/pdf/PoojaSaigal_PhD_Thesis.pdf)

Application of Generalized Inverses on Solving Fuzzy Linear Systems

Vera Miler Jerkovic
vera.miler@etf.rs
University of Novi Sad, Faculty of Technical Sciences,
Serbia

THE topic of this thesis is the presentation of the original method for solving fuzzy linear systems (FLS) using generalized inverses of a matrix. Development of science and technology has motivated investigation of methods for solving fuzzy linear systems, which parameters are rather represented by fuzzy numbers than numbers. Buckley and Qu observed the fuzzy linear system in the form of $\tilde{A}\tilde{X} = \tilde{Y}$, at the end of the last century. Further, Friedman et al. proposed a method for solving a squared FLS, in the form of $A\tilde{X} = \tilde{Y}$, which matrix A is a matrix of real coefficient and $\tilde{X}$ and $\tilde{Y}$ are fuzzy numbers vectors, while $\tilde{X}$ is unknown. Moore and Penrose presented generalized inverses of a matrix, in the middle of the last century. The most popular generalized inverses are $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{1^k\}$ and $\{5^k\}$ - inverse. They are used individually or in the combination with each other. The most applicable generalized inverse is the Moore-Penrose inverse of a matrix, which is defined as a unique solution of the system of four matrix equations. The goal of this thesis is to present the method which formulates a necessary and sufficient condition for the existence of solutions of fuzzy linear systems and gives the exact algebraic form of any solution. In addition, an efficient algorithm for determination all solutions of fuzzy linear systems is presented. In this thesis fuzzy linear systems, in the form of $A\tilde{X} = \tilde{Y}$ where real matrix A can be dimension of $m \times n$ or $n \times n$ and singular or regular, are solved. In the purpose of solving FLS where the real matrix A is mxn, the new, original method is based on generalized inverse - the Moore-Penrose inverse of a matrix. Especially, this method uses generalized $\{1, 3\}$-inverse or $\{1, 4\}$-inverse when the arbitrary, real coefficient matrix of FLS is the full rank matrix by columns or rows. The efficient algorithm for this method is presented as well as solving of the example addressed multi-criteria decision making problems. The efficient method for solving a singular, nxn fuzzy linear system, $A\tilde{X} = \tilde{Y}$, where the coefficient matrix A is a real matrix, singular or regular, using the block structure of the group inverse or any 1-inverse. Based on the presented necessary and sufficient condition for the existence of a solution, the general solution of a square FLS is obtained. Finally, infinitely many solutions of a singular FLS are presented through many interesting examples. (http://www.ftn.uns.ac.rs/539013902/disertacija)

Study of Instance Selection Methods

Alvar Arnaiz-Gonzalez

alvarag@ubu.es
Universidad de Burgos, Spain

NEW challenges have arisen for learning algorithms, as the data bases that are used for training systems have grown in size. Even a new name has been coined for referring to the problems linked to this: big data. But not only for learning algorithms, other steps of the ”Knowledge Discovery in Databases” process suffer from the same problems when the data bases’ size grows. For example, preprocessing techniques which represent one of the very first phases of KDD and their purpose is to adjust data sets to make their subsequent treatment easier.

Preprocessing methods are essential for achieving accurate models, it should never be forgotten that the quality of a trained model is strongly influenced by the quality of the data used in the training phase. One of these techniques, instance selection, is used to reduce the size of a data set by removing the instances that do not provide valuable information to the whole data set. The benefits of the instance selection methods are twofold: on the one hand, the reduction of data sets’ size makes easier the training process of different learners; on the other hand, these techniques can remove harmful instances such as noise or outliers.

This thesis focuses on the study of instance selection methods. State-of-the-art techniques were analysed and new methods were designed to cover some of the areas that had not, up until now, received the attention they deserve, more precisely they were: instance selection for regression (i) and instance selection for big data classification (ii).

Regarding to the former (i), instance selection has been extensively researched for classification but, unfortunately, not for regression. This fact can be explained because the selection of the instances in regression is much more challenging than in classification. While in the typical classification problems the membership of an instance to a class is sharply defined (an instance belongs or not to a class, and if it belongs to a class, it does not belong to the others) which facilitates the selection process, in regression there is no concept of class that can be used to guide the performing of the algorithms.

With respect to instance selection for big data classification (ii), the main drawback is the complexity of the existing methods, commonly quadratic or even higher. Instance selection has shown itself to be effective for reducing the size of the data sets while preserving their predictive capabilities. The problem that emerges at this point, is the high computational complexity that these methods have. Recently some studies have focused on it, however more scalable methods are required for instance selection with the aim of tackling the current size of data sets. In one of the chapters of the thesis, the locality sensitive hashing technique was used for designing two new instance selection algorithms of linear complexity that can be used in big data environments.

Finally, the future lines of the thesis focus on instance selection for multi-label learning. This new scenario makes the instance selection process much more challenging. (http://hdl.handle.net/10259/4830)

An Intelligent Recommender System Based on Short-term Disease Risk Prediction for Patients with Chronic Diseases in a Telehealth Environment

Raid Lafta
RaidLuaibi.Lafta@usq.edu.au
University of Southern Queensland, Australia

CLINICAL decisions are usually made based on the practitioners experiences with limited support from data-centric analytic process from medical databases. This often leads to undesirable biases, human errors and high medical costs affecting the quality of services provided to patients. Recently, the use of intelligent technologies in clinical decision making in the telehealth environment has begun to play a vital role in improving the quality of patients lives and reducing the costs and workload involved in their daily healthcare. In the telehealth environment, patients suffering from chronic diseases such as heart disease or diabetes have to take various medical tests (such as measuring blood pressure, blood sugar and blood oxygen, etc). This practice adversely affects the overall convenience and quality of their everyday living.

In this PhD thesis, an effective recommender system is proposed that utilizes a set of innovative disease risk prediction algorithms and models for short-term disease risk prediction to provide chronic disease patients with appropriate recommendations regarding the need to take a medical test on the coming day.

The input sequence of sliding windows based on the patients time series data is analyzed in both the time domain and the frequency domain. The time series medical data obtained for each chronicle disease patient is partitioned into consecutive sliding windows for analysis in both the time and the frequency domains. The available time series data are readily available in time domains which can be used for analysis without any further conversion. Yet, for data analysis in the frequency domain, Fast Fourier Transformation (FFT) and Dual-Tree Complex Wavelet Transformation (DTCWT) are applied to convert the data into the frequency domain and extract the frequency information.

In the time domain, four innovative predictive algorithms C Basic Heuristic Algorithm (BHA), Regression-Based Algorithm (RBA) and Hybrid Algorithm (HA) as well as a structural graph-based method (SG) C are proposed to study the time series data for producing recommendations. While, in the frequency domain, three predictive classifiers C Artificial Neural Network, Least Squares-Support Vector Machine, and Naive Bayes C are used to produce the recommendations. An ensemble machine learning model is utilized to combine all the used predictive models and algorithms in both the time and frequency domains to produce the final recommendation.

Two real-life telehealth datasets collected from chronic disease patients (i.e., heart disease and diabetes patients) are utilized for a comprehensive experimental evaluation in this study. The results ascertain that the proposed system is effective in analyzing time series medical data and providing accurate and reliable (very low risk) recommendations to patients suffering from chronic diseases such as heart disease and diabetes.

This research work will help provide a high-quality evidence-based intelligent decision support to clinical disease patients in significantly reducing their workload in medical checkups which otherwise have to be conducted every day in a telehealth environment. (https://drive.google.com/open?id=1Q0GrtPrCUf1ev8SdpdsvP2UIzxOpN3G-)

DATA-DRIVEN ANALYTICAL MODELS FOR IDENTIFICATION AND PREDICTION OF OPPORTUNITIES AND THREATS

Saurabh Mishra
saurabhthemishra@gmail.com
University of Maryland, United States

DURING the lifecycle of mega engineering projects such as: energy facilities, infrastructure projects, or data centers, executives in charge should take into account the potential opportunities and threats that could affect the execution of such projects. These opportunities and threats can arise from different domains; including for example: geopolitical, economic or financial, and can have an impact on different entities, such as, countries, cities or companies. The goal of this research is to provide a new approach to identify and predict opportunities and threats using large and diverse data sets, and ensemble Long-Short Term Memory (LSTM) neural network models to inform domain specific foresights. In addition to predicting the opportunities and threats, this research proposes new techniques to help decision-makers for deduction and reasoning purposes. The proposed models and results provide structured output to inform the executive decision-making process concerning large engineering projects (LEPs). This research proposes new techniques that not only provide reliable time-series predictions but uncertainty quantification to help make more informed decisions. The proposed ensemble framework consists of the following components: first, processed domain knowledge is used to extract a set of entity-domain features; second, structured learning based on Dynamic Time Warping (DTW), to learn similarity between sequences and Hierarchical Clustering Analysis (HCA), is used to determine which features are relevant for a given prediction problem; and finally, an automated decision based on the input and structured learning from the DTW-HCA is used to build a training data-set which is fed into a deep LSTM neural network for time-series predictions. A set of deeper ensemble programs are proposed such as Monte Carlo Simulations and Time Label Assignment to offer a controlled setting for assessing the impact of external shocks and a temporal alert system, respectively. The developed model can be used to inform decision makers about the set of opportunities and threats that their entities and assets face as a result of being engaged in an LEP accounting for epistemic uncertainty.

TEACHING ROBOTS WITH INTERACTIVE REINFORCEMENT LEARNING

Francisco Cruz
francisco.cruz@ucentral.cl
Universidad Central de Chile, Chile

INTELLIGENT assistive robots have recently taken their first steps toward entering domestic scenarios. It is expected that they perform tasks which are often considered rather simple for humans. However, for a robot to reach human-like performance diverse subtasks need to be accomplished in order to satisfactorily complete a given task.

An open challenging issue is the time required by a robot to autonomously learn a new task. A strategy to speed up this apprenticeship period for autonomous robots is the integration of parent-like trainers to scaffold the learning. In this regard, a trainer guides the robot to enhance the task performance in the same manner as caregivers may support infants in the accomplishment of a given task. In this dissertation, we focus on these learning approaches, specifically on interactive reinforcement learning to perform a domestic task.

First, we investigate agent-agent interactive reinforcement learning. We use an artificial agent as a parent-like trainer. The artificial agent is previously trained by autonomous reinforcement learning and afterward becomes the trainer of other agents. This interactive scenario allows us to experiment with the interplay of parameters like the probability of receiving feedback and the consistency of feedback. We show that the consistency of feedback deserves special attention since small variations on this parameter may considerably affect the learner's performance. Moreover, we introduce the concept of contextual affordances which allows reducing the state-action space by avoiding failed-states, i.e., a group of states from which it is not possible to reach the goal state. By avoiding failed-states, the learner-agent is able to collect significantly more reward. The experiments also focus on the internal representation of knowledge in trainer-agents to improve the understanding of what the properties of a good teacher are. We show that using a polymath agent, i.e., an agent with more distributed knowledge among the states, it is possible to offer better advice to learner-agents compared to specialized agents.

Thereafter, we study human-agent interactive reinforcement learning. Initially, experiments are performed with human parent-like advice using uni-modal speech guidance. We observe that an impoverished speech recognition system may still help interactive reinforcement learning agents, although not to the same extent as in the ideal case of agent-agent interaction. Afterward, we perform an experiment including audiovisual parent-like advice. The set-up takes into account the integration of multi-modal cues in order to combine them into a single piece of consistent advice for the learner-agent. Additionally, we utilize contextual affordances to modulate the advice given to the robot to avoid failed-states and to effectively speed up the learning process. Multi-modal feedback produces more confident levels of advice allowing learner-agents to benefit from this in order to obtain more reward and to gain it faster.

This dissertation contributes to knowledge in terms of studying the interplay of multi-modal interactive feedback and contextual affordances. Overall, we investigate which parameters influence the interactive reinforcement learning process and show that the apprenticeship of reinforcement learning agents can be sped up by means of interactive parent-like advice, multi-modal feedback, and affordances-driven environmental models. (http://ediss.sub.uni-hamburg.de/volltexte/2017/8609/pdf/Dissertation.pdf)

Fast, Real-Time Robot Navigation in Initially Unknown Environments via Cross-Domain Transfer Learning of Options

Olimpiya Saha
osaha@unomaha.edu
University of Nebraska at Omaha, United States

AUTONOMOUS navigation is a critical aspect of operations performed by mobile robots in numerous applications such as domestic vacuum cleaning, autonomous vehicle driving, robot-based warehouse inventory management, and, critical applications such as unmanned search and rescue, and extraterrestrial exploration. The main problem in autonomous navigation is to enable a robot to determine a collision free path between its start and goal locations while reducing the amount of energy and/or time required to move along that path, and, while satisfying constraints such as maintaining a minimum clearance with obstacles along the path. Autonomous navigation is further complicated in most real-life situations as robots sensors have limited range and the robot might not have access to an a priori or accurate map of the entire environment. Consequently, robots have to make navigation decisions based on the limited information from the environment in their immediate vicinity perceived through their sensors. Unfortunately, making decisions with limited environment information can either require time- and computationally-intensive, motion planning calculations to navigate efficiently, or, result in time- and energy-wise inefficient navigation maneuvers if the robot uses naive motion planning techniques. To address this robot navigation decision making problem in an efficient manner, we propose to use a machine learning technique called transfer learning which enables a robot to navigate efficiently in complicated environments by reusing its previous knowledge acquired from human demonstrations or through navigation in past environments. In this dissertation, we have proposed two techniques - the first technique uses a concept called experience-based learning that enables a robot to reuse learned navigation maneuvers from past environments to navigate in new environments, albeit with obstacle boundary patterns similar to those encountered in the past environments. In the second technique, we generalize this concept by relaxing the constraint that obstacle boundary patterns have to be similar and present the main technique of this dissertation called Semi-Markov Decision Processes with Uncertainty and Transfer (SMDPU-T). In the second part of this dissertation, we proposed three techniques to enhance the performance of the SMDPU-T algorithm from different aspects by utilizing inverse reinforcement learning, unsupervised learning and deep reinforcement learning. All the proposed techniques in this dissertation were implemented either on a simulated or a physical mobile, four-wheeled robot called Coroware Corobot or Turtlebot which showed that the robot using our proposed techniques could navigate successfully in new environments with previously un-encountered obstacle boundary geometries. Our experimental results on simulated robots within Webots simulator illustrate that SMDPU-T takes 24% planning time and 39% total time to solve same navigation tasks while, our hardware results on a Turtlebot robot indicate that SMDPU-T on average takes 53% planning time and 60% total time as compared to a recent, sampling-based path planner. As the final contribution of this dissertation, we extended the proposed path planning approach from a single robot to a multi-robot system with multiple ground robots, that are able to learn efficient navigation maneuvers across different environments from each others past navigation experiences through a robot cloud-like infrastructure. (https://unomaha.box.com/s/74hadlsgm4nru3a005zhlv5kaeufiphl)

Innovative Machine Learning Methods for Demand Management in Smart Grid Market

Xishun Wang
xw357@uowmail.edu.au
University of Wollongong, Australia

SMART Grid has been widely acknowledged as an efficient solution to the current energy system. Smart Grid market is a complex and dynamic market with different types of consumers and suppliers under an uncertain environment. An efficient management of Smart Grid market can benefit Smart Grid in multiple aspects, including reducing energy cost, improving energy efficiency and enhancing network reliability. This thesis focuses on improving demand management in Smart Grid market through developing innovative machine learning methods.

Firstly, this thesis studies Smart Grid market and proposes an intelligent broker model for Smart Grid market management. In the proposed broker designs, the challenges that a smart broker faces in Smart Grid market are comprehensively considered, and an adaptive and systematic model is constructed to surmount the challenges. Experimental results demonstrate that the proposed broker model can not only make much profit but also keep a good supply-demand balance. Secondly, this thesis studies how to accurately predict power demand of Smart Grid considering customer behaviors. A sparse Continuous Conditional Random Fields(sCCRF) model is proposed to explore customer behaviors. A load forecasting method through learning customer behaviors (LF-LCB) is proposed to effectively predict the demand of Smart Grid. Generally, learning customer behaviors to aggregate customers can assist decision makings towards various customers in a complex market environment. Thirdly, thesis studies effective renewable energy prediction methods through deep learning. A Deep Regression model for Sequential Data (DeepRSD) is proposed for renewable energy prediction. An alternative dropout is also proposed to effectively improve the generalization of DeepRSD. DeepRSD shows two major advantages over other known methods. 1) DeepRSD can simultaneously represent step features and temporal information. 2) DeepRSD has a strong nonlinear presentation capacity to achieve a good performance without feature engineering. Fourthly, thesis investigates state-of-the-art time-series prediction models and proposes a new effective model for time-series prediction, applying to demand prediction in Smart Grid market. The proposed model is Sparse Gaussian Conditional Random Fields (SGCRF) on top of Recurrent Neural Networks (RNN), short as CoR. CoR integrates the advantages of RNN and SGCRF and shows excellent performance in demand prediction. CoR can effectively make use of temporal correlations, nonlinearities and structured information in time-series prediction. With sufficient experiments and analysis, this thesis concludes that CoR can be a new effective model for time-series prediction in Smart Grid and broad domains. In summary, this thesis proposes several effective machine learning methods to ameliorate demand management in Smart Grid market. The proposed machine learning methods not only contribute to effective demand management of Smart Grid market in practice, but also contribute to machine learning research, as they can be applied to broad domains.

Discovery of high quality knowledge for clinical decision support system by applying semantic web technology

Seyedjamal Zolhavarieh
zolhavarieh@yahoo.com
Auckland University of Tehnology, New Zealand

WHILE the discovery of new clinical knowledge is always a good thing, it can lead to difficulties. Health experts are required to actively ensure they are informed about the latest accurate knowledge in their field. Many health experts already have access to Clinical Decision Support Systems (CDSSs). These systems aid health experts in making decisions by providing clinical knowledge. CDSS is helpful, but often has issues with the quality of knowledge extracted from knowledge sources (KSs) for decision making. Discovery of high quality clinical knowledge to support decision making is difficult. This issue is partly due to the enormous amount of research, guideline data and other knowledge published every year. Available KSs (e.g PubMed, Google scholar) are very diverse in terms of formats, structure, and vocabulary. Clinical knowledge may need to be extracted from these diverse locations and sources. To facilitate this task, many health experts focus on developing methods to manage and analyze clinical knowledge in this changeable environment. Most of KSs suffer from a lack of proper mechanism for identifying high quality knowledge. For example the PubMed search engine does not fully check some important knowledge quality metrics (QMs) such as citation, structure, accuracy and relevancy. This research has potential to make decisions easier, save time, and in turn allows the CDSSs operate more effectively. The objective of this research is to propose a knowledge quality assessment (KQA) approach to discover the high quality clinical knowledge needed for the purpose of decision making. Semantic Web (SW) technology has been used in the approach to assess how qualified knowledge is about given query. The candidate knowledge QMs were identified from related work to improve assessment of knowledge quality in CDSSs. By running a survey, the candidate knowledge QMs were reviewed and rated by health experts. Based on the survey results the knowledge QM measurements were proposed. While at an elementary stage and considered to be a proof of concept, this research offers fresh insights into what the world of healthcare will look like when knowledge quality assessment mechanism for knowledge acquisition of CDSSs is fully implemented. (http://aut.researchgateway.ac.nz/handle/10292/10966)

IEEE Computer Society
1730 Massachusetts Ave, NW
Washington, D.C. 20036-1903