

Big Data and Causality

Hossein Hassani^{1,2}, Xu Huang³, and Mansi Ghodsi²

¹hassani.stat@gmail.com

²Institute for International Energy Studies, Iran

³Faculty of Business and Law, De Montfort University, UK

Abstract

Causality analysis continues to remain one of the fundamental research questions and the ultimate objective for a tremendous amount of scientific studies. In line with the rapid progress of science and technology, the age of Big Data has significantly influenced the causality analysis on various disciplines especially for the last decade due to the fact that the complexity and difficulty on identifying causality among Big Data has dramatically increased. Data Mining, the process of uncovering hidden information from Big Data is now an important tool for causality analysis, and has been extensively exploited by scholars around the world. The primary aim of this paper is to provide a concise review of the causality analysis in Big Data. To this end the paper reviews recent significant applications of Data Mining techniques in causality analysis covering a substantial quantity of research to date, presented in chronological order with an overview table of Data Mining applications in causality analysis domain as a reference directory.

Keywords: Big Data; Data Mining Techniques; Causality Analysis.

1 Introduction

Alongside the fruits of continuous advancements of technology and information science rapidly spread across the world, the growing and accumulating information has led to the age of Big Data. Every aspect of sciences is awash with more information than ever before while the information is overflowing with a faster speed [1]. The era of Big Data challenges the approach of data analysis and decision making on many communities, from governments and e-commerce to health organizations, it furthermore led to the significant impacts on the economy, science and society at broader scale [2,4]. Data Mining itself is a relatively new and rapidly evolving subject over the last two decades, it represents the process of uncovering hidden information from Big Data (more details of the historical and modern definitions of Data Mining are referred to [5]). The popularity in the application of many Data Mining techniques are further influenced by the increasing availability of Big Data, and its ease of use for people who lack data analysis skills and statistical knowledge [6].

Causality analysis has been extensively explored by researchers during the past decades on a broad range of subjects [7–12]. It is driven by the instinctive desire of knowledge and has been considered one of the fundamental studies regardless of the research area in a broad sense. With the

advance of science and technology, the developments on causality analysis have also been overwhelmingly influenced by the age of Big Data. In order to obtain more precise and accurate extractions, successful adoption of Data Mining technique or a combination of techniques will be the crucial key for causality analysis studies with the emergence of Big Data nowadays. According to our study, the main related Data Mining techniques for causality analysis include Entity Extraction, Clustering, Association Rule Mining, Classification Techniques like Decision Trees, Neural Networks, Support Vector Machines and Naive Bayes Rule. Note that a theoretical summary of related Data Mining techniques have been provided in [13], therefore it is not reproduced here in this paper.

The aim of this paper is to provide a concise review of the Data Mining applications in causality analysis over the years in the age of Big Data¹. Prospectively, this paper also aims to summarize the rate of progress of Data Mining in causality analysis and encourage more future research to obtain much broader applications and better understandings of causal relationship regardless of the subject. In order to enable such use, the review has been organized so that interested parties could easily refer to this article alone to apprise themselves on the research that has already been conducted to date, and the resulting outcomes which have been attained. It is worthy to be highlighted that this paper not only seeks to capture majority of the significant Data Mining applications in causality analysis by classifying these based on different types of techniques, but also categorizes the specific subjects that have been exploited so the broader interested parties may find it highly beneficial to achieving a forward-looking research agenda. Moreover, the review also includes in tabular format a summary of Data Mining applications in causality analysis which can act as a quick reference guide for researchers.

The remainder of this review paper is organized such that the review of the applications of Data Mining for causality analysis is presented in Section 2 in chronological order with more specific explanations on the implementation of Data Mining techniques. The paper concludes in Section 3.

2 Data Mining Applications in Causality Analysis

In this section, the summary of the Data Mining applications in causality analysis is presented by types of Data Mining techniques and mainly following the chronological order. Additionally, Table 1 is provided as a reference directory in order to provide clear view of all reviewed literature. This table summarizes information based on the Data Mining technique(s) used and provides information relating to key techniques and software, research subject and purpose of the underlying applications².

¹Note that this paper focuses on data mining applications in causality analysis only regardless of subjects. It is formed based on a specific aspect of view, therefore it is not comparable with any other reviews of data mining applications. More relevant details, please refer to [14] that focus on time series, [15] for pharmacogenomics, [13] for crime studies, [16] for health informatics, [17] for causality analysis in biomedical informatics, [18] for fraud detection studies, etc.

²Note that an application that implemented multiple Data Mining techniques will be categorized into the review subsection of the corresponding technique that was most significantly employed.

Table 1: Overview of Data Mining applications in causality analysis.

Data Mining Techniques	Overview Summary			
	References	Key Techniques or Softwares	Specific Tested Subjects&Regions	Purpose and Function
Entity Extraction	[22–25, 27, 29, 31–42, 48–54, 56–59, 61]	lexico-syntactic patterns discovery, ambiguous patterns ranking by semantic constraints, Cause Effect Association (CEA)-based feature, distributional similarity methods, discourse relation prediction by Ruby-based discourse extraction system [62], event causality test (ECT), Penn Discourse Treebank (PDTB [55], GENIA Event corpus [63], NAGGNER [64], ProNormz [65], FrameNet [66], Learning Based Java modeling language (LBJ) [67], Optical Character Recognition (OCR), Map-Reduce framework, LinkedData [26], TimeML [60], CATENA [61]	Text mining in computational linguistics [22, 23], French text mining [31], medical database text mining [39], MSN search engine queries from temporal logs [40], open domain text with classification between encoding and not encoding causation [24], event relation extraction in Biomedical Information [41, 42, 48, 49, 51–53], event causality in open domain text (CNN [56], NYT [25], TimeBank [58, 59], MEDLINE [42], TempEval-3 [61]), verbal events [27], event causality by focusing on verb-noun pairs [29], Japanese text mining [34–36], German multilog [37], Arabic text mining [32, 33, 38], Spanish [54]	Extract valuable causal relation information from unstructured text data regardless of subjects (even data by different languages), identifying causal relationships as significant implementation of various natural language processing (NLP) applications, identify and extract causal events from temporal query logs.
Cluster Analysis	[70–76, 78–81, 83–87]	Attribute-Oriented Induction (AOI) technique [88, 89], M-Correlator [90], K-Means Clustering, Hierarchical Clustering Technique, Dirichlet process clustering [91], Mahalanobis distance [92], Topographic Mapping of Proximity data (TMP) [93], Louvain method [77], DBSCAN (density-based clustering) [82], TRACCLUS [94]	Information/Cyber Security [78], cardiac arrhythmia [70], diabetes and hemophilus influenza B (HiB) vaccine [71], crime prevention [79], gene expression [72], financial development and economic growth [80], fMRI datasets (neuroimaging study) [73–76], taxi trajectory data in Beijing [81], industrial operation management [84, 85], natural calamities [86, 87], air quality and meteorological data in Beijing [83]	Grouping low-level data and emphasising hidden significant relationships, discover causal relations in terms of the real life large data sets or observational medical trails data, comprehensive causal analysis in terms of each cluster, recovering community network structure based on interactive behavior between different nodes in biomedical image processing application [73–76].
Association Rule	[98–106, 108–114, 116–122]	Apriori algorithm [96], TETRAD Software [124], LCD algorithm [99], DBMiner [125], CU-path algorithm [100], inter transaction mining [126], profit mining [127]	Market basket data [100], discrete data [101], stock market in Taiwan [103–106], retrospective cohort study in health, gene expression and phenotypes [121], associated drug reaction for drug safety study [110–114, 116–118, 122], medical and social research [107–109, 120]	Discover potential causal rules in observational data, identify causal relationships on real, large-scale data set, mine the causal relationship between drugs and their associated adverse drug reactions.
Classification Techniques	[135–140, 142–148, 155, 155–163, 165–171, 176–181]	CART, C4.5 Algorithm, J48 in Weka [182], tree kernel method [183], naive algorithm, MinEntropy [142], Bayesian neural network [156], Tree Augmented Naive Bayes Classifiers [184], K-dependence Bayesian classifier, recurrent neural network model, LIBSVM [185].	medical outcome [155], adverse drug reaction [156], the Louisiana and Helgoland wether database [136, 137], Stock Market Monitoring from PDA [139], failure diagnosis [142], Intrusion Detection System by KDD’99 data set [178], question answering [138], air transport safety (Netherlands) [144], Detecting Network Neutrality violations [143], gene regulatory network [157–160], gene-gene and gene-environment interaction [166], cancer diagnosis in breast cancer study [180, 181], cancer subgroup mining with heterogeneous treatment causal effects [148], identify semantic relations in text [165, 176], fMRI data [168], genome-wide causal variants study [167], triggering relation discovery on cyber security [170], clinical diagnose and treatment [161, 169], stock market in Shanghai [140], Spanish mining accident [146], industrial occupational safety [171], the Titanic data set, the adult data set census income and 5 groups of synthetic data set [147], SemEval-2010-Task8 dataset [177]	fast algorithm to discover causal signals in large-scale data set especially when the target or outcome variable is fixed, mining and selecting optimal parameters for further causal analysis modelling, identify the causes of failures in large Internet sites, demonstrate human volitionally regulate hemodynamic signals from circumscribed regions of the brain leading to area-specific behavioral consequences, identify genetic variants associated with disease, determine classification model to help on obtaining efficient decision for treating cancer patients.

2.1 Entity Extraction

Entity Extraction is a process that identifies particular patterns such as text, images, or audio materials depending on the availability of extensive amounts of clean input data [4, 19, 20]. As concluded in [21], the main approaches for entity extraction are: lexical-lookup, rule-based, statistic-based and machine learning.

The generally applicable lexico-syntactic patterns that refer to the causal relationship are detected by the text mining technique in [22, 23], in which they firstly identify lexico-syntactic patterns that can express the causal relationship, then the ambiguous patterns acquired are validated and ranked by semantic constraints on nouns and verbs.

A text mining system combining the classification technique is proposed to analyse the open domain text for detecting causations between a verb phrase and a subordinate clause [24]. In which, the authors firstly classified 1270 sentences from the TREC5 corpus and detect 170 marked and explicit causation for forming the syntactic patterns. The experiments by SemCor 2.1 corpus then identified 1068 instances and 517 causations by matching syntactic patterns, which yielded a high performance of averaging over 0.9 of recall ratio.

In [25], the authors generated the Pundit system to perform causal reasoning in textually represented unrestricted environments. In which, a large information source spanning more than 150 years was achieved from New York Times by optical character recognition (OCR), and the entity graph was obtained by Map-Reduced framework according to the relations between the concepts by LinkedData cloud project [26]. By proposing a framework that automatically harvests a network of causal-effect terms from a large web corpus, the authors in [50] used a data driven approach to solve the problem of commonsense causality reasoning between short texts.

By focusing on the causality detection of the verbal events, a knowledge-based approach was developed to evaluate the prediction performance of the verb pairs so to identify the causality relationships by employing the knowledge-rich metrics [27]. More specifically, 12,000 documents from the English Gigaword corpus and 3,000 articles on news (also used by [28]) were collected and analyzed with desired performance. In order to improve the performance further without just relying on shallow linguistic features, the authors employed additional types of knowledge on semantic classes along with linguistic features, which achieved about 30% improvement in accuracy for the causality recognition by focusing on the verb-noun pairs [29].

The causality analysis by text mining of different languages other than English has also been widely exploited by scholars. By referring to Force Dynamics [30], an automatic tool named COATIS was presented to identify causality links in French text data by targeting linguistic indicators of causality in sentences [31]. The Arabic Discourse Treebank was firstly generated in [32] and evolved in [33] along with the first algorithm to identify the discourse connectives and causal relations in Arabic text. The authors in [34–36] exploited semantic relation, context and association features in Japanese, in which, innovatively, the proposed semantic relation features also contained the ones that have less obvious relations to causality. The authors in [37] focused on the causal discourse relations in transcripts of spoken multilog in German, where a linguistically-motivated, rule-based

annotation system was proposed. Additionally, the authors in [38] studied on the extraction of causal relations in Arabic using linguistic patterns, in which the authors achieved a precision of 78%.

The authors in [39] studied about causal knowledge extraction especially on the medical data. In which, the information explicitly expressed in medical abstracts in the Medline database was explored. Over 200 abstracts were analyzed as training sample covering four different medical areas, and 68 patterns were constructed for the 35 causality identifiers. The medical linguistic markers of causal expressions were then identified, extracted and analyzed by pattern matching based on the syntactic parse trees of sentences.

In terms of search engine query logs, the causal relations indicate the causation and effect link between two queries. The authors exploited the MSN search engine data in [40] and developed a 2-dimensional visualization tool to present the causal relationships. In which, events are firstly identified by efficient statistical frequency threshold; the causal relations of queries are then mined by geometric features of the events; by combining the Granger causality test, the causal relations are finally re-ranked based on the test coefficients. Their experiments obtained accurate and effective performance of detecting the events in temporal query logs and causal relations of queries.

The causal relation analysis in terms of biomedical information has also been explored by many scholars by extracting the targeted entities. Among which, a relation extraction method based on named entity-driven information extraction was proposed in [41] for discovering the causal relations in the BioNLP'09 task. By focusing on the mining challenge of protein-protein interactions, a web-based text mining tool (named PPInterFinder) was implemented in [42] to extract causal relations with promising performance of 66% accuracy on five standard corpora (AIMED [43], BioInfer [44], HPRD50 [45], IEPA [46] and LLL [47]). An annotation scheme BioCause was defined for enriching biomedical domain corpora with causal relations in [48]. Furthermore, the BioCause corpus was upgraded in [49] by adopting a self-learning algorithm considering command relations in parse trees and positional features, which improved the performances of identifying causal relations in biomedical scientific discourse. The authors in [51] extracted hidden information from biomedical literature to reveal the Protein-Protein, Drug-Drug causal interactions. In terms of the biological expression language, Track 4 at BioCreative V was presented in [52] to identify and extract various levels of information so to achieve the extraction of causal networks from text, whilst the authors in [53] described the new corpora that succeed to capture causal relationships not only between proteins or chemicals, but also complex events such as biological processes or disease states. In [54], both named entity recognition and event extraction were adopted to detect the causality relationship between drugs and diseases through the electronic health record in Spanish.

The event causality identification was studied by combining discourse relation predictions (through the Penn Discourse Treebank (PDTB) [55]) and distributional similarity methods in a global inference procedure [56]. Their experiments on the collected articles from CNN has proved additional improvements towards determining event causality through text mining. Furthermore, the authors in [57] extracted the causal events from the cause-effect pairs identified, and built an abstract causality network with effective performances on identifying high-level causality rules behind specific causal events.

The authors in [58,59] conducted developments based on TimeLM [60] and presented a framework for identifying causal signals and causal relations between events. Furthermore, a sieve-based system CATENA was introduced in [61] to extract causal relation from English natural language text with promising state-of-the-art performances proved by both TempEval-3 and TimeBank-Dense data.

2.2 Cluster Analysis

Cluster analysis indicates grouping objects into categories/clusters based only on information found in the data which describes the objects and their relationships, such that the objects in a group will be similar (or related) to one another [68]. Note that a comprehensive survey of cluster analysis algorithms is provided in [69].

A generalization of the classical cluster analysis is proposed in [70] to contribute on identifying certain structure of causality. Specifically, the approach proposed is a subsequent cluster analysis applied to the centers of clusters obtained in the first clustering, and it was experimented on the data of cardiac arrhythmia with promising performance.

More applications of cluster analysis technique have then been conducted on medical studies like diabetes, gene expression, neuroimaging, etc. The authors in [71] studied on the causal relation between type 1 diabetes and the Hemophilus influenza B (HiB) vaccine by adopting clustering technique on a large clinical trial data. It was proved that exposure to HiB immunization is associated with an increased risk of type 1 diabetes. Functional clustering on genes was conducted based on the similar expression patterns as well as the causality analysis in [72], which outperformed the usual approach and provides better understanding of gene expression data sets as well as their regulatory networks. A novel aspect of causal analysis on neuroimaging data sets was presented in [73], which adopted the concept of informative clustering so to group the variables from different brain regions in terms of their shared information on the future of another targeted variable.

Moreover, clustering technique has also been widely adopted for the analyses of fMRI data set. A cluster Granger causality method was proposed in [74] to analyze the connectivity between regions of interests based on fMRI data set. The clusters of voxels were defined to prepare the multidimensional series for further causality analysis, and the experimental results showed promising performance on detecting interregional connections. A pair-wise clustering approach was proposed in [75] to be applied on the large scale Granger Causality Index interactions from processing the fMRI data. It proved with promising performance on reconstructing the structure of the original network and better understanding the interactions between different nodes of the network. Furthermore, the authors introduced the non-linear mutual connectivity analysis framework in [76]. The non-metric network clustering technique was adopted based on the Louvain method [77] to recover the network structure so to contribute on the investigation of causal relations between regions of the motor cortex.

In order to prevent the overwhelming working loads of information security (INFOSEC) system and to maintain the proper responses in a timely fashion, clustering techniques are adopted in [78] to group low-level alert data into high-level aggregated alerts based on the information of corresponding attributes so to conduct causal analysis on the INFOSEC problem regarding security alert correlation and relationships among attacks.

Considering data mining approach as a proactive decision-support tool in terms of crime prevention, the authors in [79] proposed a framework of uncovering hidden-causal-effect knowledge and reveal the shift around effect by studying the temporal crime activity data from National Police Agency in Taiwan. Clustering mining technique was firstly implied for mining the significant parameters, then a rule extraction algorithm based on association rule mining technique was employed to discover causal relations.

In terms of the financial development and economic growth studies, clustering analysis technique was combined with a regime switching panel vector auto regression model to identify directional effects in finance-growth causality based on a sample of 69 countries [80], in which the clustering analysis identified the presence of convergence clubs based on data properties and the results confirmed the growth to financial development unidirectional causality and coexistence of bidirectional causality for most countries.

By exploring the trajectory data collected from taxis in Beijing, the authors in [81] adopted density-based clustering method DBSCAN [82] to better discovery of the region structure and capturing causal relationship among regions. More specifically, the causal time-varying dynamic Bayesian network was applied to reveal the evolution of their causal time-varying structures. Then the density-based clustering assisted to directly identify regions for a particular space-time interval from trajectories and further analyse the spatio-temporal behavior of drivers driving from one region to another selected based on the causal structures. Another study that focusing on urban big data of Beijing in [83] combined K clusters technique, pattern mining and bayesian learning to investigate the spatiotemporal causal structure between air quality and meteorological data.

Focusing on the causality extraction of the wear of machinery, the authors in [84] adopted the lubricating oil analysis data to investigate the causal rules for wear conditions of the equipments by combining the cluster analysis technique, which makes it possible to have more detailed diagnosis regarding wear conditions of machinery in the future tasks. Another industrial application in [85] targeted the study of causal mining for large-scale complex industrial plant, the authors combined the dynamic time warping-based K-means clustering method and modified group Granger causality to detect the root cause of disturbances that may impact the over all control performance and lead to inferior quality.

Another development of clustering method in [86] is the cluster sequence mining technique that extracts pattern from numerical multidimensional event sequences. Specifically, it extracts patterns with a pair of clusters that satisfies space proximity of the individual clusters and time proximity in time intervals between events from different clusters. It was further adopted to the causality analysis of earthquake occurrences based on an earthquake event sequence in Japan after 2011. Similarly, the authors in [87] studied the cause-and-effect relationships between hydrological parameters and the Majiagou landslide movement in China with data mining techniques including primarily two-step cluster analysis as well as association rule mining.

2.3 Association Rule

Association Rule is a technique for investigating the possibility of simultaneous occurrence of data [95], it aims to mine all rules in the database that satisfy some minimum support and minimum confidence constraints [97]. It was initially proposed in [96] as a method of discovering interesting co-occurrences in supermarket data. Note that its implementation for large data sets can indicate the strength of association among data attributes, which can be further examined for causal relations [98].

The local causal discovery (LCD) algorithm was firstly proposed in [99] by focusing on the observational data. It was illustrated underlying constraint-based algorithm that combined Association Rule Mining technique to contribute on the causal discovery.

Silverstein et al. [100] focused on the research of market basket data and proposed the novel algorithms by combining Association Rule Mining so to determine causal relationships on large scale data sets. The experimental results on both census data and text data indicated sufficient performance on identifying causal structures with feasible computation time. Another experiment with contraceptive method choice data in Indonesia was conducted in [101] by a comprehensive comparison study between causal Association Rule Mining and Bayesian Network method.

The authors in [102] proposed a model of mining causality among multi-value variables based on partitioning, which is a generalization of both item-based and quantitative Association Rule Mining. It was proved to establish on extracting causal rules with reduced unnecessary information in large databases.

By focusing on the stock market in Taiwan, Hsieh et al. [103] applied inter-transaction Association Rule Mining to identify the causal relation between upstream and downstream companies, which is significantly beneficial information for investors. The authors further extended the research on profit mining model in [104, 105] so to better satisfy the investors' expectation on causal relations discovery regardless of the format of the knowledge. The latest research in [106] specifically focused on the closed item sets and developed the new version of profit mining approach with more efficient performance.

Aiming at more efficiently identifying potential causal relationship in observational data, the authors in [108, 109] adopted Association Rule Mining together with cohort studies to develop the approach of causal Association Rule Mining. The proposed approach has been evaluated on 24 synthetic data sets and 8 frequently used public data sets of medical and social research with comparison of the Bayesian network method with stable and efficient performance.

The causal Association Rule Mining frameworks were proposed in [110, 111] and applied to mine potential causal associations in electronic patient health database where the drug-related events of interests occur infrequently. Experimental results showed promising performance and effectively identified causal relations in the database. On the other hand, Association Rule Mining was applied in [112–114] to identify causal signals for drug and adverse reaction based on the user contributed content data in social media. More details can be found in [115], in which the authors reviewed data mining techniques that have been studied in the area of drug safety to identify signals of adverse drug reactions from various data sources. According to the data from the United States Food and

Drug Administration adverse events reporting system, the authors in [116] applied Association Rule Mining to identify drug cause-and-effect interactions. A recent research in [117] studied the electronic patient database and presented a temporal association mining approach to effectively identify the cause-and-effect relationships between two events within a patient case based on the occurrences of various symptoms, so to prevent the serious consequences of drug-drug interactions. Note that a recent review of drug-drug interactions through data mining techniques can be found in [118] for more details.

A general approach to discover causal relations in large observational databases of binary variables was proposed in [119], in which the partial associations were also taken into account so to conduct better and more efficient performance on identifying causal relations with combined cause variables. Similarly, another data-driven application is [120], in which, the authors analyzed a dataset consisting of 2200 incidents of military activity surrounding ISIS and the forces that oppose it in the Islamic State by adopting logic programming and association rule mining. The authors discovered causal relationships between terrorist activity and military operations as well as rules indicating fire, suicide attacks, etc.

The application is also exploited for gene expression data in [121]. In which, the authors proposed the dynamic association rule algorithm that will help to efficiently select a subset of significant genes for subsequent analysis of the causal relationships between genes and phenotypes. The experiments were conducted on the analyses of for microarray datasets and one next generation sequencing dataset, which all shows efficient and accurate performances on identifying influential genes of a disease.

Yadav et al. [122] further improved the Association Rule Mining on causality detection from observational data by adopting the Rubin-Neyman causal model [123]. The authors evaluated the proposed causal rule mining framework that transition from Association Rule Mining towards causal inference in subpopulation on the electronic health records data and proved sufficiently performance on extracting the controversial findings of the causal effect of a class of cholesterol drugs on type two diabetes.

2.4 Classification Techniques

As one of the most fundamental and significant Data Mining techniques, classification is defined as the task of assigning objects to one of several predefined categories [68] and discovering a small set of rules in the database to form an accurate classifier [128]. It contains a few specific types of techniques including Decision Trees, Neural Networks, Support Vector Machines, Naive Bayes Rule, etc.

In terms of applying classification techniques in Data Mining for causality analysis, many implementations combined more than one specific type of classification technique. Therefore, the review which follows is classified by each significant technique in chronological order and depending on circumstances, those complex combination cases are not reproduced.

Decision Trees

Decision Trees([129–131]) are applied to accomplish the classification task by giving a series of carefully crafted questions about the attributes of the test record [68]. When an answer is achieved, it will be followed by a question until the category of this attribute is concluded. All the series of questions and possible answers are carefully predefined, as well as the process repeated to all subsets of the tree³. Many algorithms have been developed for getting the most reasonably optimal Decision Tree with good accuracy in a timely manner. For example, CART [129], C4.5 [130], ID3 [131], Hunt’s Algorithm [132], SLIQ [133], and SPRINT [134].

A tool named TimeSleuth was proposed for discovering causality in [135], in which C4.5 decision tree algorithm was adopted along with time involved into the input as preprocessing step and adjusting accordingly based on the original temporal relations as a post processing step. Furthermore, the authors developed TIMERS method in [136] and exploited on the Louisiana wether database and the Helgoland weather database, in particular, the proposed method was based on finding classification rules to predict the value of a decision attribute using various of observed condition attributed. The later second version of TIMERS algorithm was introduced in [137] to classify the relationship between a decision attribute and a number of condition attributes as instantaneous, causal, or acausal(possibly containing hidden common causes), which was the latest development for the upgraded algorithm.

By expanding the Decision Trees technique into language processing and computational linguistics, Girju [138] proposed an automatic detection of causal relations framework for question answering, in which the C4.5 decision tree algorithm was adopted for extracting the lexical and semantic constraints referring to causation in English text.

Classification techniques like Decision Trees has also been widely exploited in finance studies. Kargupta et al. [139] proposed an experimental mobile data mining system named MobiMine that adopted decision tree mining technique to facilitate the monitoring process by identifying the interesting behaving stocks and detecting their causal relationship with different features characterizing the stocks. Moreover, by studying 13 years data from the Shanghai Stock Exchanges, observational data-based causal analysis was applied on stock predictions in [140]. In particular, the authors applied the CART algorithm (decision tree mining) and proposed the causal feature selection algorithm to select more representative features for better stock prediction modeling. Additionally, note that a review of data mining techniques in financial application can be found in [141].

By focusing on the failure diagnosis in large Internet sites, Chen et al. [142] presented a decision tree learning approach to identify the causes of failures. Actual failure data from eBay was evaluated and the presented algorithm successfully identified 13 out of 14 true causes of failure. Similarly, a system named NANO was proposed in [143] for detecting Network Neutrality violation, in particular, a decision tree based classification method was employed to infer the discrimination criteria.

The applications were further extended to the air traffic system. The authors in [144] studied on the air traffic system of the Netherlands international airport Schiphol and proposed the operational causal model based on the decision tree technique for finding causes of incidents and accidents

³It is possible that we can get many different Decision Trees from the same given set of cases. The final choice depends on the research and the individual circumstances.

therefore to quantification of the probability of adverse events in the aviation industry. The authors further developed the causal model for air transport safety in [145] by linking event sequence diagrams, fault-trees and Bayesian belief nets to form a homogeneous mathematical model.

The authors in [146] focused on the Spanish mining accidents studies and adopted decision trees and bayesian classifiers among other data mining techniques to analyze the main causes of mining accidents based on a database composed of almost 70,000 occupational accidents and fatality reports corresponding to the decade 2003-2012 in the Spanish mining sector. The study successfully concluded a few causal rules that can significantly develop suitable prevention policies to reduce mining accidents.

A recent theoretical advancement was presented in [147], in which the authors proposed a causal decision tree model based on the improvement of normal decision tree mining technique. The causal relations were interpreted by the nodes with a compact graphical representation of all uncovered causal relationships, additionally, the calculation was efficient and the performances were promising based on results of three sets of experiments (including the Titanic data set, the adult data set census income and 5 groups of synthetic data set). Furthermore, the authors in [148] proposed the survival causal tree method to mine patient subgroups with heterogeneous treatment causal effects from censored observational data. It was applied to identify cancer subtypes at molecular level, which can be significantly helpful to select the most suitable treatment for individual patients comparing to the clinical diagnoses, which can lead to better survival chances and less suffering due to inaccurate diagnoses.

Neural Networks (NN)

Neural Networks is one of the most important tools for classification that achieved effective and successful performances in many real world classification tasks [154]. Recent research has established convincing evidence to this end showing that Neural Networks has high tolerance to noisy data and the ability to classify untrained patterns. According to [149], Neural Networks is able to estimate the posterior probabilities, which provides the basis for establishing classification rule and performing statistical analysis [149–153].

The applications that adopting Neural Networks technique with causality analysis are found mainly focusing on medical and genetic studies. Tu [155] evaluated the Neural Networks and logistic regression approaches on the medical outcome studies, in particular, Neural Networks required less formal statistical training, showed impressive ability to implicitly detect complex nonlinear relationships between dependent and independent variables, and interactions between predictor variables. A data mining system based on a Bayesian neural network was presented in [156] to assist on minimizing the limitations of the current system and highlight strong causal relations between specific drugs and corresponding adverse drug reactions based on the WHO database. In terms of the reconstruction of gene regulatory network, Wahde and Hertz [157] adopted the recurrent Neural Network model and applied on a set of actual expression data from the development of rat central nervous system. Furthermore, the authors in [158–160] applied the same model on the gene expression data to capture the nonlinear dynamics of gene networks and reveal genetic regulatory interactions from

expression profiles. A recent research in [161], the authors proposed the paradigm causal phenotype discovery by combining the Neural Networks data mining technique and pairwise log likelihood non-Gaussian structural causal inference model. It was aimed to discover latent representations of illness that are causally predictive and a series of phenotype experiments have been applied to a few clinical time series data collected during the delivery of care in intensive care units at large hospitals.

Another extension of Neural Networks technique has been exploited on conflict analysis, the authors in [162] developed and tested a neural network model of Cold War interstate conflicts and evaluated its performance on data from 1885 to 1992. The experiment revealed the extent to which the Cold War causal structure was representative of earlier historical contexts.

Moreover, a theoretical development was proposed by the authors in [163], in which the Neural Networks approach was adopted as a bridge between model-free and model-based causality detection approaches to better recognize dynamics in complex data sets and identify causal flows occurring in a system of time series. In particular, the approach required no priori assumptions and had been proved sufficient approach by simulations; by adopting the non-uniform embedding, it was capable of providing the optimal path of mapping between input and output spaces; it also led to a further development with respect to traditional Granger causality approaches when redundant variables are involved.

Support Vector Machines (SVM)

Support Vector Machines (SVM) is a method of separating two classes using an optimal separating hyperplane which minimizes the classification error and it has been widely applied as an significant classification technique for Data Mining, pattern recognition, etc [164]. Moreover, SVM was also extensively employed in cooperation of causality analysis on various subjects.

The SVM technique was employed for text processing in [165], in which it assisted on the automatic identification of a set of seven semantic relations between nominals in English sentences. Specifically, the approach adopted various sets of lexical, syntactic and semantic features extracted from various knowledge sources and achieved an accuracy of 76.3% on the SemEval 2007 task.

In terms of genetic studies, it was adopted in [166] to identify and characterize high order gene-gene and gene-environment interactions, which indicated significant importance of understanding the underlying biological mechanisms of complex diseases and the complex relationships that control the process. By focusing on the genome-wide association studies, the authors in [167] combined the SVM technique to rank the causal variants and associated regions with promising performance by real data set experiments.

Furthermore, Lee et al. [168] applied SVM technique in medical study on the fMRI data to observe changes in the spatial activation patterns in the brain across the training sessions, so to be prepared for the implementation of multivariate Granger causality modelling to calculate directed causal influences between spatially distributed voxels of the brain. The authors in [169] also focused on the medical studies aspect and conducted the study to extract causality patterns for the problem-action relations in discharge summaries so to present a chronological view of a patients' problem and an doctor's action. In which, the causal relationship between events from clinical narratives

are investigated and the clinical semantic unit is classified by adopting SVM. Experiment has been applied on Korean discharge summaries with about 80% of accurate performance on effectively classifying clinical problem-action relations.

By focusing on the cyber security studies, Zhang et al. [170] proposed a traffic analysis method to reason the occurrences of network event and target the stealthy malware activities in order to efficiently discover the underlying triggering relations of a massive amount of network events and detect malware activities on a host. In particular, three different classifiers (SVM, Naive Bayes, Bayesian network) were compared on training and classifying the data based on chosen features. Similarly, a study that focused on the occupational safety of a steel plant in India was conducted by combining the SVM data mining technique in [171]. In which, the SVM served as sufficient classifier to identify causal rules and provided 88% accuracy of predicting the accidents based on a database comprising almost 5000 occupational accidents reports from an integrated steel plant from 2010 to 2012.

Naive Bayes Rule

Naive Bayes classifier is proposed in [172] and uses Bayes Rule to compute the probability of each class given the instance, assuming the attributes are conditionally independent given the label [173]. In general, it is simple and easy to understand, convenient for implementation, and one of the most efficient and effective inductive learning algorithms for machine learning and Data Mining [174].

In terms of the cause-effect relations in natural language text, Chang and Choi [175] worked on extracting causal relations that exist between two events expressed by noun phrases or sentences, in order to do so, the lexical pair probability and the cue phrase probability were introduced along with the employment of the Naive Bayes classifier. Experiments on data sets from LA TIMES and Wall Street Journal were conducted with promising performance on causal relation extraction. Sorgente et al. [176] used the Naive Bayes Rule mining technique to identify cause-effect pairs based on the dependency relations between the words. Evaluations were obtained on the SemEval-2010(Task 8) data set with encourage results of over 70% precision score achieved. A restricted hidden Naive Bayes model was proposed in [177] for text mining and extracting event causal relations from text. A new category feature of causal connectives were included to classify among candidates of causal pair, and the proposed approach has been proved to be able to cope with the possible interactions between features so to improve the causality extraction performance.

The Naive Bayes Rule has also been adopted for the cyber security studies in [178], the authors compared the Naive Bayes Rule with Decision Trees on the performance of intrusion detection by using KDD'99 data set. Furthermore, they applied the Naive Bayes Rule approach on the DARPA2000 data set to reduce the high volume of reported alerts and detect complex and coordinated attacks in [179].

As one of the significant areas that Data Mining techniques have been widely exploited on, the Naive Bayes Rule technique was adopted on breast cancer clinical data set in [180] to discover cause-specific death classes and propose a graphical structure of key attributes describing the conditional dependency among attributes. This contributed on extracting the causal relationships among clinical

variables and therefor allowed the more efficient and accurate cancer diagnosis for treating cancer patients. Furthermore, the authors in [181] also focused on the causal relationship identification among clinical variables for breast cancer and adopted Naive Bayes classifier along with improved flexible k-dependence network through target-based-encoding for numerization of categorical values with the assistance of target class. The results are further improved in diagnosing cancer causing attributes, even for extremely strong positive relationships.

3 Conclusion

This paper is driven by the importance of Data Mining technique in the age of Big Data in terms of the ultimate research subject of causality analysis in a broader horizon. Given the vast amount of research on Data Mining developments and applications, it indicates an emerging demand on the better understanding of both Big Data and causality analysis regardless of subject. Following a thorough research we are able to present a list of Data Mining techniques as the most frequently adopted at present for causality analysis of Big Data. These include Entity Extraction, Clustering, Association Rule Mining, Classification Techniques like Decision Trees, Neural Networks, Support Vector Machines and Naive Bayes Rule.

Not only to provide a review of tremendous amount of applications, this paper also achieves to obtain a directory table with categories by Data Mining techniques and details of research subjects and objects, therefore the broader interested parties can easily refer to this article alone to apprise themselves on the up to date conducted research. Moreover, this paper also contributes on directing researchers from various areas to achieve a forward-looking research agenda. In general, this paper has the advance in building the bridge between two significant groups of researchers: one is the scholars who expertise in Big Data analysis and Data Mining techniques seeking for more applications; and the other group is the researchers who are interested in the causality analysis in particular areas containing Big Data and also enthusiast in adopting frontier techniques. Table 1 in particularly functions as a useful resource or 'quick guide' which summarises the Data Mining applications in causality analysis while providing useful information on not only on the software and purpose, but also the research subject that have been exploited.

This paper has captured considerably adequate amount of recent significant applications. Considering the wide disciplines that have been exploited and the amount of applications that were captured, the classification techniques are the most popular form of Data Mining in terms of causality analysis. This also in line with the findings from Data Mining in crime analysis [13]. Moreover, we notice that the Data Mining techniques are seldom used for mining multivariate causality analysis model in economics, whereas in text processing, cyber security, biomedical information and medical study these methods are extremely popular and well exploited. This further highlights the disciplinary differences that exist across subjects and the emerging need of popularizing the use of Data Mining technique in the age of Big Data. The review of a tremendous amount of successful applications has provided the future insight and immeasurable potentials of Data Mining technique

in causality analysis. It is genuinely expected that this review paper can contribute on better understanding of the causality analysis with Big Data regardless of subjects and promoting further advancements of Data Mining techniques as well as their broader applications in causality analysis.

References

- [1] Mayer-Schonberger, V., and Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt. New York.
- [2] Chen, H., Chiang, R. H., and Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly* 36(4): 1165–1188.
- [3] Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., and Chau, M. (2004). Crime data mining: a general framework and some examples. *Computer* 37(4): 50–56.
- [4] Gupta, G. K. (2006). *Introduction to Data Mining with case studies*. PHI Learning Pvt. Ltd.
- [5] Hassani, H., Saporta, G., and Silva, E. S. (2014). Data Mining and Official Statistics: The past, the present and the future. *Big Data* 2(1): 34–43.
- [6] Fayyad, U., and Uthurusamy, R. (2002). Evolving data into mining solutions for insights. *Communications of the ACM* 45(8): 28–31.
- [7] Granger, C. W. (1988). Some recent development in a concept of causality. *Journal of Econometrics* 39(1-2): 199–211.
- [8] Soytaş, U., and Sari, R. (2003). Energy consumption and GDP: causality relationship in G-7 countries and emerging markets. *Energy Economics* 25(1): 33–37.
- [9] Hassani, H., Zhigljavsky, A., Patterson, K., and Soofi, A. (2010). A comprehensive causality test based on the singular spectrum analysis. *Causality in Science*: 379–406.
- [10] Sugihara, G., May, R., Ye, H., Hsieh, C. H., Deyle, E., Fogarty, M., and Munch, S. (2012). Detecting causality in complex ecosystems. *Science* 338(6106) 496–500.
- [11] Hassani, H., Huang, X., Gupta, R., and Ghodsi, M. (2016). Does sunspot numbers cause global temperatures? A reconsideration using non-parametric causality tests. *Physica A: Statistical Mechanics and its Applications* 460: 54–65.
- [12] Ghodsi, Z., Huang, X., and Hassani, H. (2017). Causality analysis detects the regulatory role of maternal effect genes in the early Drosophila embryo. *Genomics Data* 11: 20–38.
- [13] Hassani, H., Huang, X., Silva, E. S., and Ghodsi, M. (2016). A review of data mining applications in crime. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 9(3): 139–154.
- [14] Fu, T. C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence* 24(1): 164–181.
- [15] Hahn, U., Cohen, K. B., Garten, Y., and Shah, N. H. (2012). Mining the pharmacogenomics literature: a survey of the state of the art. *Briefings in Bioinformatics* 13(4): 460–494.

- [16] Herland, M., Khoshgoftaar, T. M., and Wald, R. (2014). A review of data mining using big data in health informatics. *Journal of Big Data* 1(1): 2.
- [17] Kleinberg, S. and Hripcsak, G. (2011). A review of causal inference for biomedical informatics. *Journal of Biomedical Informatics* 44(6): 1102–1112.
- [18] Sharma, A. and Panigrahi, P. K. (2012). A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. *International Journal of Computer Applications* 39(1): 37–47.
- [19] Cowie, J., and Lehnert, W. (1996). Information extraction. *Commun. ACM* (39)1: 80–91.
- [20] Chinchor, N. A. (1998). Overview of MUC-7/MET-2. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, April 1998.
- [21] Chau, M., Xu, J. J., and Chen, H. (2002). Extracting meaningful entities from police narrative reports. In: *Proceedings of the 2002 Annual National Conference on Digital Government Research*: 1–5.
- [22] Girju, R., and Moldovan, D. I. (2002). Text mining for causal relations. In *FLAIRS Conference*: 360–364.
- [23] Girju, R., and Moldovan, D. (2002). Mining answers for causation questions. In *AAAI symposium on mining answers from texts and knowledge bases*.
- [24] Blanco, E., Castell, N., and Moldovan, D. I. (2008). Causal Relation Extraction. In *Lrec*.
- [25] Radinsky, K., Davidovich, S., and Markovitch, S. (2012). Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*: 909–918. ACM.
- [26] Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *Semantic services, interoperability and web applications: emerging concepts*: 205–227.
- [27] Riaz, M., and Girju, R. (2013). Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*.
- [28] Riaz, M., and Girju, R. (2010). Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference*: 361–368. IEEE.
- [29] Riaz, M., and Girju, R. (2014). Recognizing causality in verb-noun pairs via noun and verb semantics. *EACL*: 48.
- [30] Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science* 12(1): 49–100.

- [31] Garcia, D. (1997). COATIS, an NLP system to locate expressions of actions connected by causality links. In *International Conference on Knowledge Engineering and Knowledge Management*: 347–352. Springer Berlin Heidelberg.
- [32] Al-Saif, A., and Markert, K. (2010). The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. In *LREC*.
- [33] Alsaif, A., and Markert, K. (2011). Modelling discourse relations for Arabic. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*: 736–747. Association for Computational Linguistics.
- [34] Hashimoto, C., Torisawa, K., De Saeger, S., Oh, J. H., and Kazama, J. I. (2012). Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*: 619–630. Association for Computational Linguistics.
- [35] Hashimoto, C., Torisawa, K., Kloetzer, J., Sano, M., Varga, I., Oh, J. H., and Kidawara, Y. (2014). Toward Future Scenario Generation: Extracting Event Causality Exploiting Semantic Relation, Context, and Association Features. In *ACL (1)*: 987–997.
- [36] Hashimoto, C., Torisawa, K., Kloetzer, J., and Oh, J.H. (2015). Generating Event Causality Hypotheses through Semantic Relations. In *AAAI*: 2396–2403.
- [37] Bögel, T., Hautli-Janisz, A., Sulger, S., and Butt, M. (2014). Automatic detection of causal relations in German multilogs. In *14th Conference of the European Chapter of the Association for Computational Linguistics*: 20–27.
- [38] Sadek, J., and Meziane, F. (2016). Extracting arabic causal relations using linguistic patterns. *ACM Transactions on Asian and Low-Resource Language Information Processing* 15(3): 14.
- [39] Khoo, C. S., Chan, S., and Niu, Y. (2000). Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*: 336–343. Association for Computational Linguistics.
- [40] Sun, Y., Xie, K., Liu, N., Yan, S., Zhang, B., and Chen, Z. (2007). Causal relation of queries from temporal logs. In *Proceedings of the 16th international conference on World Wide Web*: 1141–1142. ACM.
- [41] Pyysalo, S., Ohta, T., Kim, J. D., and Tsujii, J. I. (2009). Static relations: a piece in the biomedical information extraction puzzle. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*: 1–9. Association for Computational Linguistics.
- [42] Raja, K., Subramani, S., and Natarajan, J. (2013). PPInterFindera mining tool for extracting causal relations on human proteins from literature. *Database*: bas052.

- [43] Bunesu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K., and Wong, Y. W. (2005). Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine* 33(2): 139–155.
- [44] Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Jarvinen, J., and Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics* 8(1): 50.
- [45] Fundel, K., Kffner, R., and Zimmer, R. (2007). RelExRelation extraction using dependency parse trees. *Bioinformatics*, 23(3), 365-371.
- [46] Ding, J., Berleant, D., Nettleton, D., and Wurtele, E. (2002). Mining MEDLINE: abstracts, sentences, or phrases. In *Proceedings of the pacific symposium on biocomputing 7*: 326–337.
- [47] Nédellec, C. (2005). Learning language in logic-genic interaction extraction challenge. In *Proceedings of the 4th Learning Language in Logic Workshop (LLL05) 7*: 1–7.
- [48] Mihăilă, C., Ohta, T., Pyysalo, S., and Ananiadou, S. (2013). BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics* 14(1): 2.
- [49] Mihăilă, C., and Ananiadou, S. (2014). Semi-supervised learning of causal relations in biomedical scientific discourse. *Biomedical Engineering Online* 13(2): S1.
- [50] Luo, Z., Sha, Y., Zhu, K. Q., Hwang, S. W., and Wang, Z. (2016, March). Commonsense Causal Reasoning between Short Texts. In *KR*: (pp. 421-431).
- [51] Mahendran, D. and Nawarathna, R. D. (2016). An automated method to extract information in the biomedical literature about interactions between drugs. In *Advances in ICT for Emerging Regions (ICTer), 2016 Sixteenth International Conference on* (pp. 155-161). IEEE.
- [52] Rinaldi, F., Ellendorff, T. R., Madan, S., Clematide, S., van der Lek, A., Mevissen, T., and Fluck, J. (2016). BioCreative V track 4: a shared task for the extraction of causal network information using the Biological Expression Language. *Database*: baw067.
- [53] Fluck, J., Madan, S., Ansari, S., Kodamullil, A. T., Karki, R., Rastegar-Mojarad, M., Catlett, N.L., Hayes, W., Szostak, J., Hoeng, J. and Peitsch, M. (2016). Training and evaluation corpora for the extraction of causal relationships encoded in biological expression language (BEL). *Database*: baw113.
- [54] Casillas, A., Pérez, A., Oronoz, M., Gojenola, K., and Santiso, S. (2016). Learning to extract adverse drug reaction events from electronic health records in Spanish. *Expert Systems with Applications* 61: 235–245.
- [55] Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., and Webber, B. L. (2007). The penn discourse treebank 2.0 annotation manual. *IRCS Technical Reports Series*: 203.

- [56] Do, Q. X., Chan, Y. S., and Roth, D. (2011). Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*: 294–303. Association for Computational Linguistics.
- [57] Zhao, S., Wang, Q., Massung, S., Qin, B., Liu, T., Wang, B., and Zhai, C. (2017). Constructing and embedding abstract event causality networks from text snippets. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*: (pp. 335-344). ACM.
- [58] Mirza, P., and Tonelli, S. (2014). An Analysis of Causality between Events and its Relation to Temporal Information. In *COLING*: 2097–2106.
- [59] Mirza, P. (2014). Extracting Temporal and Causal Relations between Events. In *ACL (Student Research Workshop)*: 10–17.
- [60] Pustejovsky, J., Lee, K., Bunt, H., and Romary, L. (2010). ISO-TimeML: An International Standard for Semantic Annotation. In *LREC 10*: 394–397.
- [61] Mirza, P., and Tonelli, S. (2016). CATENA: CAusal and TEmporal relation extraction from NATural language texts. In *The 26th International Conference on Computational Linguistics*: 64–75.
- [62] Lin, Z., Ng, H. T., and Kan, M. Y. (2014). A PDTB-styled end-to-end discourse parser. *Natural Language Engineering* 20(02): 151–184.
- [63] Kim, J. D., Ohta, T., and Tsujii, J. I. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics* 9(1): 10.
- [64] Kalpana, R., Suresh, S., and Jeyakumar, N. (2012). NAGGNERa hybrid named entity tagger for tagging human proteins/genes. In *Proceedings of the tenth Asia Pacific Bioinformatics Conference*. Melbourne, Australia.
- [65] Suresh, S., Kalpana, R., and Jeyakumar, N. (2011). ProNormzan automated web server for human proteins and protein kinases normalization. In *Proceedings of the second International Conference on Bioinformatics and Systems Biology (INCOBS)*. Chidambaram, India.
- [66] Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., and Scheffczyk, J. (2006). FrameNet II: Extended theory and practice.
- [67] Rizzolo, N., and Roth, D. (2010). Learning Based Java for Rapid Development of NLP Systems. In *LREC (5)*: 313–323.
- [68] Pang-Ning, T., Steinbach, M., and Kumar, V. (2006). Introduction to data mining. In *Library of Congress*.
- [69] Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science* 2(2): 165–193.

- [70] Matuszewski, A. (2002). Double Clustering: A Data Mining Methodology for Discovery of Causality. In *Intelligent Information Systems*: 227–236. Physica-Verlag HD.
- [71] Classen, J. B. and Classen, D. C. (2002). Clustering of cases of insulin dependent diabetes (IDDM) occurring three years after hemophilus influenza B (HiB) immunization support causal relationship between immunization and IDDM. *Autoimmunity* 35(4): 247–253.
- [72] Fujita, A., Severino, P., Kojima, K., Sato, J. R., Patriota, A. G., and Miyano, S. (2012). Functional clustering of time series gene expression data by Granger causality. *BMC Systems Biology* 6(1): 137.
- [73] Wu, G., Liao, W., Stramaglia, S., Chen, H., and Marinazzo, D. (2013). Recovering directed networks in neuroimaging datasets using partially conditioned Granger causality. *Brain Connectivity* 3(3): 294–301.
- [74] Sato, J. R., Fujita, A., Cardoso, E. F., Thomaz, C. E., Brammer, M. J., and Amaro, E. (2010). Analyzing the connectivity between regions of interest: an approach based on cluster Granger causality for fMRI data analysis. *Neuroimage* 52(4): 1444–1455.
- [75] Wismüller, A., Nagarajan, M. B., Witte, H., Pester, B., and Leistritz, L. (2014). Pair-wise clustering of large scale Granger causality index matrices for revealing communities. In *SPIE Medical Imaging*: 90381R-90381R. International Society for Optics and Photonics.
- [76] Wismüller, A., Wang, X., DSouza, A. M., and Nagarajan, M. B. (2014). A Framework for Exploring Non-Linear Functional Connectivity and Causality in the Human Brain: Mutual Connectivity Analysis (MCA) of Resting-State Functional MRI with Convergent Cross-Mapping and Non-Metric Clustering. *arXiv preprint arXiv:1407.3809*.
- [77] Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Louvain method: finding communities in large networks.
- [78] Qin, X., and Lee, W. (2003). Statistical causality analysis of infosec alert data. In *International Workshop on Recent Advances in Intrusion Detection*: 73–93. Springer Berlin Heidelberg.
- [79] Li, S. T., Kuo, S. C., and Tsai, F. C. (2010). An intelligent decision-support model using FSOM and rule extraction for crime prevention. *Expert Systems with Applications* 37(10), 7108–7119.
- [80] Chow, W. W., and Fung, M. K. (2013). Financial development and growth: a clustering and causality analysis. *The Journal of International Trade & Economic Development* 22(3): 430–453.
- [81] Wong, R. K., Chu, V., Ghanavati, M., and Hamzehei, A. (2015). Trajectory analysis based on clustering and casual structures. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- [82] Birant, D., and Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatialtemporal data. *Data & Knowledge Engineering* 60(1): 208–221.

- [83] Zhu, J. Y., Zhang, C., Zhi, S., Li, V. O., Han, J., and Zheng, Y. (2016). p-Causality: Identifying Spatiotemporal Causal Pathways for Air Pollutants with Urban Big Data. *arXiv preprint arXiv:1610.07045*.
- [84] Ide, D., Ruike, A., and Kimura, M. (2015). Extraction of causalities and rules involved in wear of machinery from lubricating oil analysis data. In *The Second International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC2015)* (p. 16).
- [85] Yuan, T., Li, G., Zhang, Z., and Qin, S. J. (2016). Deep causal mining for plant-wide oscillations with multilevel granger causality analysis. In *American Control Conference (ACC)*: 5056–5061. IEEE.
- [86] Okada, Y., Fukui, K. I., Moriyama, K., and Numao, M. (2015). Cluster Sequence Mining: Causal Inference with Time and Space Proximity Under Uncertainty. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 293–304). Springer International Publishing.
- [87] Ma, J., Tang, H., Hu, X., Bobet, A., Zhang, M., Zhu, T., Song, Y. and Eldin, M. A. E. (2017). Identification of causal factors for the Majiagou landslide using modern data mining methods. *Landslides* 14(1): 311–322.
- [88] Cai, Y. (1989). Attribute-oriented induction in relational databases. *Doctoral dissertation*. Simon Fraser University.
- [89] Han, J., Cai, Y., and Cercone, N. (1993). Data-driven discovery of quantitative rules in relational databases. *IEEE transactions on Knowledge and Data Engineering* 5(1): 29–40.
- [90] Porras, P. A., Fong, M. W., and Valdes, A. (2002). A mission-impact-based approach to INFOSEC alarm correlation. In *International Workshop on Recent Advances in Intrusion Detection*: 95–114. Springer Berlin Heidelberg.
- [91] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2004). Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. In *NIPS*: 1385–1392.
- [92] De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems* 50(1): 1–18.
- [93] Bishop, C. M., Svensen, M., and Williams, C. K. (1998). GTM: The generative topographic mapping. *Neural Computation* 10(1): 215–234.
- [94] Lee, J. G., Han, J., and Whang, K. Y. (2007). Trajectory clustering: a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*: 593–604. ACM.
- [95] Yun, H., Ha, D., Hwang, B., and Ryu, K. H. (2003). Mining Association Rules on significant rare data using relative support. *Journal of Systems and Software* 67(3): 181–191.

- [96] Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining Association Rules between sets of items in large databases. *In: ACM SIGMOD Record 22*: 207–216.
- [97] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. *In Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).
- [98] Mazlack, L. J. (2008). Considering Causality In Data Mining. *In International Conference on Software Engineering*.
- [99] Cooper, G. F. (1997). A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery 1(2)*: 203–224.
- [100] Silverstein, C., Brin, S., Motwani, R., and Ullman, J. (2000). Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery 4(2-3)*: 163–192.
- [101] Bowes, J., Neufeld, E., Greer, J. E., and Cooke, J. (2000). A comparison of association rule discovery and Bayesian network causal inference algorithms to discover relationships in discrete data. *In Conference of the Canadian Society for Computational Studies of Intelligence*: 326–336. Springer Berlin Heidelberg.
- [102] Zhang, S., and Zhang, C. (2002). Discovering causality in large databases. *Applied Artificial Intelligence 16(5)*: 333–358.
- [103] Hsieh, Y. L., Yang, D. L., and Wu, J. (2005). Using data mining to study upstream and downstream causal relationship in stock market. *Computer 1*: F02.
- [104] Hsieh, Y. L., Yang, D. L., and Hsu, F. R. (2012). An effective mining algorithm for profit mining. *In Computer, Consumer and Control (IS3C), 2012 International Symposium*: 106–110. IEEE.
- [105] Hsieh, Y. L., Yang, D. L., and Wu, J. (2014). Effective application of improved profit-mining algorithm for the interday trading model. *The Scientific World Journal*: ID874825.
- [106] Hsieh, Y. L., Yang, D. L., Wu, J., and Chen, Y. C. (2016). Efficient Mining of Profit Rules from Closed Inter-Transaction Itemsets. *Journal of Information Science and Engineering 32(3)*: 575–595.
- [107] Li, J., Liu, L., & Le, T. (2015). *Practical approaches to causal relationship exploration*. Springer.
- [108] Li, J., Le, T. D., Liu, L., Liu, J., Jin, Z., and Sun, B. (2013). Mining causal association rules. *In Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference*: 114–123. IEEE.
- [109] Li, J., Le, T. D., Liu, L., Liu, J., Jin, Z., Sun, B., and Ma, S. (2016). From observational studies to causal rule mining. *ACM Transactions on Intelligent Systems and Technology (TIST) 7(2)*: 14.

- [110] Ji, Y., Ying, H., Dews, P., Mansour, A., Tran, J., Miller, R. E., and Massanari, R. M. (2011). A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. *IEEE Transactions on Information Technology in Biomedicine* 15(3): 428–437.
- [111] Ji, Y., Ying, H., Tran, J., Dews, P., Mansour, A., and Massanari, R. M. (2013). A method for mining infrequent causal associations and its application in finding adverse drug reaction signal pairs. *IEEE transactions on Knowledge and Data Engineering* 25(4): 721–733.
- [112] Yang, C. C., Yang, H., Jiang, L., and Zhang, M. (2012). Social media mining for drug safety signal detection. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*: 33–40. ACM.
- [113] Yang, C. C., Yang, H., and Jiang, L. (2014). Postmarketing drug safety surveillance using publicly available health-consumer-contributed content in social media. *ACM Transactions on Management Information Systems (TMIS)* 5(1): 2.
- [114] Yang, H., and Yang, C. C. (2015). Using health-consumer-contributed data to detect adverse drug reactions by association mining with temporal analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)* 6(4): 55.
- [115] Karimi, S., Wang, C., Metke-Jimenez, A., Gaire, R., and Paris, C. (2015). Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys (CSUR)* 47(4): 56.
- [116] Ibrahim, H., Saad, A., Abdo, A., and Eldin, A. S. (2016). Mining association patterns of drug-interactions using post marketing FDAs spontaneous reporting data. *Journal of Biomedical Informatics* 60: 294–308.
- [117] Ji, Y., Ying, H., Tran, J., Dews, P., Lau, S. Y., and Massanari, R. M. (2016). A functional temporal association mining approach for screening potential drugdrug interactions from electronic patient databases. *Informatics for Health and Social Care* 41(4): 387–404.
- [118] Vilar, S., Friedman, C., and Hripcsak, G. (2017). Detection of drugdrug interactions through data mining studies using clinical sources, scientific literature and social media. *Briefings in Bioinformatics*: bbx010.
- [119] Jin, Z., Li, J., Liu, L., Le, T. D., Sun, B., and Wang, R. (2012). Discovery of causal rules using partial association. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on* (pp. 309–318). IEEE.
- [120] Stanton, A., Thart, A., Jain, A., Vyas, P., Chatterjee, A., and Shakarian, P. (2015). Mining for Causal Relationships: A Data-Driven Study of the Islamic State. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 2137–2146. ACM.

- [121] Chen, S. C., Tsai, T. H., Chung, C. H., and Li, W. H. (2015). Dynamic association rules for gene expression data analysis. *BMC Genomics* 16(1): 786.
- [122] Yadav, P., Prunelli, L., Hoff, A., Steinbach, M., Westra, B., Kumar, V., and Simon, G. (2016). Causal Inference in Observational Data Causal Inference in Observational Data. *arXiv preprint arXiv:1611.04660*.
- [123] Sekhon, J. S. (2008). The Neyman-Rubin model of causal inference and estimation via matching methods. *The Oxford Handbook of Political Methodology*: 2.
- [124] Scheines, R., Spirtes, P., Glymour, C., and Meek, C. (1994). TETRAD II: Users manual and software.
- [125] Han, J., Fu, Y., Wang, W., Chiang, J., Gong, W., Koperski, K., ... and Xia, B. (1996). DBMiner: A System for Mining Knowledge in Large Relational Databases. In *KDD 96*: 250–255.
- [126] Tung, A. K. H., Lu, H., Han, J., and Feng, L. (2003). Efficient mining of intertransaction association rules. *IEEE Transactions on Knowledge and Data Engineering* 15(1): 43–56.
- [127] Wang, K., Zhou, S., and Han, J. (2002). Profit mining: From patterns to actions. In *International Conference on Extending Database Technology*: 70–87. Springer Berlin Heidelberg.
- [128] Liu, B., Hsu, W. and Ma, Y. (1998). Integrating classification and association rule mining. In *Proceedings of the 4th*.
- [129] Breiman, L, Friedman, J., Olshen, R. and Stone C. (1984). Classification and regression trees. *Belmont: Wadsworth*.
- [130] Quinlan, J. R. (1992). C4.5: program for machine learning. *Morgan Kaufmann*.
- [131] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106.
- [132] Hunt, J. W. and Szymanski, T. G. (1977). A fast algorithm for computing longest common subsequences. *Communications of the ACM*, 20(5), 350–353.
- [133] Mehta, M., Agrawal, R. and Rissanen, J. (1996). SLIQ: A fast scalable classifier for data mining. In *Advances in Database TechnologyEDBT’96 (pp. 18–32)*. Springer Berlin Heidelberg.
- [134] Shafer, J.C., Agrawal, R. and Mehta, M. (1996). "SPRINT: A Scalable Parallel Classifier for Data Mining", *Proc. of the 22th Int’l Conference on Very Large Databases, Mumbai (Bombay), India, Sept.*
- [135] Karimi, K., and Hamilton, H. J. (2002). TimeSleuth: A tool for discovering causal and temporal rules. In *Tools with Artificial Intelligence, 2002.(ICTAI 2002). Proceedings. 14th IEEE International Conference on*: 375–380. IEEE.

- [136] Karimi, K., and Hamilton, H. J. (2003). Distinguishing causal and acausal temporal relations. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*: 234–240. Springer Berlin Heidelberg.
- [137] Hamilton, H. J., and Karimi, K. (2005). The TIMERS II algorithm for the discovery of causality. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*: 744–750. Springer Berlin Heidelberg.
- [138] Girju, R. (2003). Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering 12*: 76–83. Association for Computational Linguistics.
- [139] Kargupta, H., Park, B. H., Pittie, S., Liu, L., Kushraj, D., and Sarkar, K. (2002). MobiMine: Monitoring the stock market from a PDA. *ACM SIGKDD Explorations Newsletter* 3(2): 37–46.
- [140] Zhang, X., Hu, Y., Xie, K., Wang, S., Ngai, E. W. T., and Liu, M. (2014). A causal feature selection algorithm for stock prediction modeling. *Neurocomputing* 142: 48–59.
- [141] Zhang, D., and Zhou, L. (2004). Discovering golden nuggets: data mining in financial application. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(4), 513–522.
- [142] Chen, M., Zheng, A. X., Lloyd, J., Jordan, M. I., and Brewer, E. (2004). Failure diagnosis using decision trees. In *Autonomic Computing Proceedings*: 36–43. IEEE.
- [143] Tariq, M. B., Motiwala, M., Feamster, N., and Ammar, M. (2009). Detecting network neutrality violations with causal inference. In *Proceedings of the 5th international conference on Emerging networking experiments and technologies* (pp. 289–300). ACM.
- [144] Ale, B. J. M., Bellamy, L. J., Cooke, R. M., Goossens, L. H. J., Hale, A. R., Roelen, A. L. C., and Smith, E. (2006). Towards a causal model for air transport safetyan ongoing research project. *Safety Science* 44(8): 657–673.
- [145] Ale, B. J., Bellamy, L. J., Van der Boom, R., Cooper, J., Cooke, R. M., Goossens, L. H., ... and Spouge, J. (2009). Further development of a Causal model for Air Transport Safety (CATS): Building the mathematical heart. *Reliability Engineering & System Safety* 94(9): 1433–1441.
- [146] Sanmiquel, L., Rossell, J. M., and Vintro, C. (2015). Study of Spanish mining accidents using data mining techniques. *Safety Science* 75: 49–55.
- [147] Li, J., Ma, S., Le, T., Liu, L., and Liu, J. (2016). Causal Decision Trees. *IEEE Transactions on Knowledge and Data Engineering*.
- [148] Zhang, W., Le, T. D., Liu, L., Zhou, Z. H., and Li, J. (2017). Mining heterogeneous causal effects for personalized cancer treatment. *Bioinformatics*: btx174.

- [149] Richard, M. D. and Lippmann, R. P. (1991). Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation*, 3(4), 461–483.
- [150] Zhang, G. P. (2000). Neural networks for classification: a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews*, IEEE Transactions on, 30(4), 451–462.
- [151] Gish, H. (1990). A probabilistic approach to the understanding and training of neural network classifiers. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on (pp. 1361–1364)*. IEEE.
- [152] Shoemaker, P. A. (1991). A note on least-squares learning procedures and classification by neural network models. *Neural Networks*, IEEE Transactions on, 2(1), 158–160.
- [153] Wan, E. A. (1989). Neural network classification: a Bayesian interpretation. *IEEE transactions on neural networks/a publication of the IEEE Neural Networks Council*, 1(4), 303–305.
- [154] Widrow, B., Rumelhart, D. E. and Lehr, M. A. (1994). Neural networks: Applications in industry, business and science. *Communications of the ACM*, 37(3), 93–105.
- [155] Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology* 49(11): 1225–1231.
- [156] Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., and De Freitas, R. M. (1998). A Bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology* 54(4), 315–321.
- [157] Wahde, M., and Hertz, J. (2000). Coarse-grained reverse engineering of genetic regulatory networks. *Biosystems* 55(1): 129–136.
- [158] Vohradský, J. (2001). Neural network model of gene expression. *The FASEB Journal* 15(3): 846–854.
- [159] Xu, R., Venayagamoorthy, G. K., and Wunsch, D. C. (2007). Modeling of gene regulatory networks with hybrid differential evolution and particle swarm optimization. *Neural Networks* 20(8): 917–927.
- [160] Noman, N., Palafox, L., and Iba, H. (2013). Reconstruction of gene regulatory networks from gene expression data using decoupled recurrent neural network model. In *Natural Computing and Beyond*: 93–103. Springer Japan.
- [161] Kale, D. C., Che, Z., Bahadori, M. T., Li, W., Liu, Y., and Wetzels, R. (2015). Causal phenotype discovery via deep networks. In *AMIA Annual Symposium Proceedings*: 677. American Medical Informatics Association.

- [162] Lagazio, M., and Russett, B. (2003). A neural network analysis of militarized disputes, 1885-1992: Temporal stability and causal complexity. *Temporal Stability and Causal Complexity*: 28–62. University of Michigan Press, New Jersey.
- [163] Montalto, A., Stramaglia, S., Faes, L., Tessitore, G., Prevete, R., and Marinazzo, D. (2015). Neural networks with non-uniform embedding and explicit validation phase to assess Granger causality. *Neural Networks* 71: 159–171.
- [164] Suykens, J. A., and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters* 9(3): 293–300.
- [165] Beamer, B., Bhat, S., Chee, B., Fister, A., Rozovskaya, A., and Girju, R. (2007). UIUC: A knowledge-rich approach to identifying semantic relations between nominals. In Proceedings of the 4th International Workshop on Semantic Evaluations (pp. 386-389). Association for Computational Linguistics.
- [166] Chen, S. H., Sun, J., Dimitrov, L., Turner, A. R., Adams, T. S., Meyers, D. A., ... and Hsu, F. C. (2008). A support vector machine approach for detecting gene-gene interaction. *Genetic Epidemiology* 32(2): 152–167.
- [167] Roshan, U., Chikkagoudar, S., Wei, Z., Wang, K., and Hakonarson, H. (2011). Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. *Nucleic Acids Research* 39(9): e62.
- [168] Lee, S., Ruiz, S., Caria, A., Veit, R., Birbaumer, N., and Sitaram, R. (2011). Detection of cerebral reorganization induced by real-time fMRI feedback training of insula activation a multivariate investigation. *Neurorehabilitation and Neural Repair* 25(3): 259–267.
- [169] Seol, J. W., Yi, W., Choi, J., and Lee, K. S. (2017). Causality patterns and machine learning for the extraction of problem-action relations in discharge summaries. *International Journal of Medical Informatics* 98: 1–12.
- [170] Zhang, H., Yao, D. D., and Ramakrishnan, N. (2014). Detection of stealthy malware activities with traffic causality and scalable triggering relation discovery. In Proceedings of the 9th ACM symposium on Information, computer and communications security (pp. 39-50). ACM.
- [171] Sarkar, S., Vinay, S., Pateshwari, V., and Maiti, J. (2016). Study of optimized SVM for incident prediction of a steel plant in India. In *India Conference (INDICON) IEEE Annual*: 1–6. IEEE.
- [172] Langley, P., Iba, W. and Thompson, K. (1992). An analysis of Bayesian classifiers. In *AAAI* 90: 223–228.
- [173] Kohavi, R. (1996). Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In *KDD* (pp. 202–207).
- [174] Zhang, H. (2004). The optimality of naive Bayes. *AA*, Vol. 1(2), 3.

- [175] Chang, D. S., and Choi, K. S. (2004). Causal relation extraction using cue phrase and lexical pair probabilities. In *International Conference on Natural Language Processing* (pp. 61-70). Springer Berlin Heidelberg.
- [176] Sorgente, A., Vettigli, G., and Mele, F. (2013). Automatic Extraction of Cause-Effect Relations in Natural Language Text. *DART AI IA*: 37–48.
- [177] Zhao, S., Liu, T., Zhao, S., Chen, Y., and Nie, J. Y. (2016). Event causality extraction based on connectives analysis. *Neurocomputing* 173: 1943–1950.
- [178] Amor, N. B., Benferhat, S., and Elouedi, Z. (2004). Naive bayes vs decision trees in intrusion detection systems. In *Proceedings of the 2004 ACM symposium on Applied computing*: 420–424. ACM.
- [179] Benferhat, S., Kenaza, T., and Mokhtari, A. (2008). A naive bayes approach for detecting coordinated attacks. In *Computer Software and Applications, 2008. COMPSAC'08. 32nd Annual IEEE International* (pp. 704-709). IEEE.
- [180] Wang, L. (2015). Mining causal relationships among clinical variables for cancer diagnosis based on Bayesian analysis. *BioData Mining* 8(1): 13.
- [181] Krishna, M. S. G. and Singh, S. (2016). Identification of causal relationships among clinical variables for cancer diagnosis using multi-tenancy. In *Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference*: pp. 1511–1516. IEEE.
- [182] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11(1): 10–18.
- [183] Collins, M., and Duffy, N. (2001). Convolution kernels for natural language. In *NIPS 14*: 625–632.
- [184] Alcobé, J. R. (2002). Incremental learning of tree augmented naive Bayes classifiers. In *Ibero-American Conference on Artificial Intelligence* (pp. 32-41). Springer Berlin Heidelberg.
- [185] Chang, C. C., and Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3): 27.