Whitepaper

Development of a Collaborative Data Collection and Automated Diagnostic Rule Generation System for Mental and Behavioral Disorders

Charles Kim, Ph.D.
Professor
Electrical Engineering and Computer Science
College of Engineering
Howard University
202-806-4821 (ckim@howard.edu)

1. Project Summary:

Mental and behavioral disorders affect a large portion of American populations and more acutely African-Americans. They are difficult diagnose accurately, and are not known of their triggering events or symptoms. The behaviors and the changes in the behaviors of the subjects have been known to shed some light in the diagnosis; however, the range and the diversity of the behaviors and changes of them are either too numerous to be practically used in clinical diagnostics or, in all practicality, simply ignored except traditionally accepted ones. Moreover, the successful diagnosis of these disorders depends on understanding not only the behavioral change of a subject but also the environments which surround the subject such as populace, culture, religion, and social dynamics. Therefore, in diagnosis of a subject in a local area, if a global spectrum of information regarding the various subjects under varieties of socio-cultural environments is provided, a local clinician would be more successful in correct diagnosis of her patient. She may now consider once-ignored symptoms, unregistered personality changes, for instance, in her diagnosis.

This kind of global confirmation of local diagnostics, by which inclusion or exclusion or ranking of certain symptomatic variables would be made possible, would be achieved by a framework of collaborative data sharing and diagnostic variable discovery and learning from the data. Open to collaborators in all areas, professions, and socio-cultural groups, the framework would become a comprehensive knowledge aggregation for diagnostics. This project aims to develop a diagnostic rule generation system with collaborative data collection for mental and socio-behavioral disorders. The system would, first, utilize collaborators-provided heterogeneous and diverse data of past and concurrent studies and observation, second, discover most relevant symptom variables pertinent to the disorders, and, third, generate diagnostic rules with probability and margin of error. The system updates the process of pattern discovery and diagnostic rule generation as new data enter. Thus, the system would enhance understanding in global perspectives in diagnosing patients of a local area or group.

The brain of the diagnostic rule generation system is an inference algorithm, which is principled around inductive reasoning with information entropy, at the maximum of which the unbiased probability is obtained, and at the minimum of which best symptom variables can be discovered to affirmation of the disorders or negation and thus the derived diagnostic rule becomes the simplest and most reliable. A few symptom variables with high probability would be selected for a trend analysis for the prediction of the occurrence of the disorder onset or epidemics by tracking the times between the occurrences of the symptomatic events. Considering the

globally located data sources, the framework of collaborative data fusion and diagnosis rule generation will be implemented in a networked server-client system. The inference algorithm would reside in a server and the data would be imported to the server via web-based client loading by the would-be-formed collaborators.

The duration of the project is 3 years. In the first year, an inference algorithm will be developed and implemented in a server computer system, and test and validation of the algorithm will be performed with readily available health assessment data such as Behavioral Risk Factor Surveillance System (BRFSS) data from Centers for Disease Control and Prevention. In the second year, a global collaborator network will be formed among clinicians and researchers for a few selected mental and behavioral disorders, and an intranet-based data collection and test-bed will be developed with multiple computers connected as a host, a server, and clients. In the third and last year, the entire year will be dedicated to evaluation of the framework for its effectiveness in helping clinicians and researchers in correct diagnosis of the disorders as well as the trend analysis and tracking of high probability symptom variables.

2. Project Description:

The proposed framework of collaborative data sharing and diagnostic variable discovery system would aggregate diagnostic knowledge and understanding for clinicians and researchers and thus enhance the correct and timely diagnosis for mental and behavioral disorders which affect a large portion of American populations and more acutely African-Americans. Mental and behavioral disorders are difficult diagnose accurately, and are not known of their triggering events or symptoms, and are difficult to cure and prevent unless detected early. The inductive inference and collaborative data collection implemented on an intranet-based network would become a comprehensive learning system for detection, treatment, prevention, and prediction of such mental and behavioral disorders.

Research Plan

This project aims to improve diagnostic accuracy of mental and socio-behavioral disorders for clinicians and researchers in dealing with regional patients by incorporating them into a framework of a global data collection and diagnostic rule generation system. The rationale of this project is in its emphasis on the advantage and benefit in diagnosis of a disorder for regional subjects, when utilized with all of the decisions made before or being made for the disorder in every possible geographical and social environment.

The diagnostic rule generation is achieved by an inference algorithm which, utilizing collaborators-provided heterogeneous and diverse sources of past and concurrent diagnoses and observations, produces highly correlational symptomatic variables of a disorder which could reveal their onset, detection, epidemics, cure, or prevention and, from and using the high variables, extracts diagnostic rules for the disorder.

The inference of diagnostic rule extraction is built around inductive reasoning with information entropy. Once all symptomatic variables are divided by the disorder, a set of rules is extracted in steps with rule patterns by the variables in the order from the highest diagnostic classification reliability (or confidence) to the lowest. The rule would produce, upon given diagnostic data from the collaborators, a probability for affirming a disorder in question for subjects. New data may be easily added to the rule extraction process and the additional data would update the rule and the probability, making it a learning decision assist system for diagnosis.

The specific goal of this project is to implement the diagnostic rule generation and collaborative data collection system in an intranet-based client-server system so that collaborators upload local or regional diagnostic data and access to the global diagnostic rule generated from the inference algorithm from the collected global data.

To achieve the goal, an intermediate objective is validation of the inference algorithm with readily available health assessment data such as Behavioral Risk Factor Surveillance System (BRFSS) data from Centers for Disease Control and Prevention. The eventual objective is, after the computer implementation of the algorithm and the client-server data collection network, to perform an extensive validation process which measures data collection effectiveness, new symptomatic variable discovery, and overall diagnostic accuracy improvement.

Research Strategy:

Mental and behavioral disorders affect a large portion of American populations and more acutely African-Americans [1]. They are difficult diagnose accurately, are not known of their triggering events or symptoms, and are difficult to cure and prevent unless detected early. The range and the diversity of the symptomatic behaviors and changes of them are numerous in clinical diagnostics and in all practicality simply ignored except traditional ones [2]. Considering additional burdens for a correct diagnosis of understanding not only the behavioral change of a subject but also the environments which surround the subject such as populace, culture, religion, and social dynamics, a global spectrum of diagnostic information for the various subjects under varieties of socio-cultural environments is essential for successful diagnosis of subjects in a local area.

There is a challenge in integrating behavioral and diagnostic data from diverse sources and then automatically adding new knowledge and understanding on the disorders and then making it available to the clinicians, researchers, and the patients. The significance of the project is that the proposed framework answers the challenge by the rule extraction and self-learning inference algorithm for diagnostic decision assist for distinctive individuals utilized with global data collected from the collaborators specialized in diverse regions, cultures, age groups, gender, and other various socio-cultural environments. The information entropy-based inference algorithm enhances knowledge and understanding of the disorders by generating diagnostic rules from the collaborated data and further updating the rules with a self-learning mechanism as more new data are entered.

Considering blurry borders of mental and behavioral disorder classes and the very possibility that a seemingly irrelevant and minor behavior or behavioral change or impairment, noticed only in non-traditional groups for example, may play a decisive role for detecting and diagnosing certain disorders, it is furthermore significant that the global data sharing and diagnostic system prevents those in disorder diagnosis for a distinctive group of individuals from narrowly relying on just the locally available symptomatic variables in the group alone. Also, the collaborative data collection framework would reduce positive false rate by increased diversity in database.

The innovation in this project is that it attempts to generate diagnostic rules automatically which are globally optimized by the discovery of the symptomatic variables which is high in probability and low in margin of error in diagnosing mental disorders for patients in any geographical and socio-cultural environments. The project incorporates global perspectives to a local distinctive individual diagnosis in determining disorders or onset of such disorders. The proposed framework would lead to comprehensive knowledge gain, early detection and identification, cure or reduction of pain, and prevention of the disorders. Subsequently, the proposed framework

would provide speedy enhancement in inclusion (or exclusion) of new and old symptomatic variables in diagnosis, and thus result in instantaneous and live updates on characterization and more accurate and reliable diagnostics of the disorders.

Approach

The strategy for this project has 3 steps. The first step builds a theoretically sound and clinician-like reasoning inference machine and tests it stand-alone with historical data of disease diagnosis. The second step builds a human network of collaborators who will to join the collaborative data collection, knowledge aggregation, global diagnostic rule generation and usage of such a rule in one's environment, and continued improvement through active participation. The second step also builds physical communication network which accommodates the collaboration activity through internet utilizing server-client networking and porting the stand-alone inference machine to the server. The third step verifies and validates the entire framework and runs the system online and real time. A specific analysis for performance measure in correct diagnosis includes the difference in the success rate of diagnosis by the global rule and that by one collaborator in a specific environment.

Following the strategy, there are three major tasks planned to develop the framework and achieve the goals and objectives: Inference machine development and initial validation, collaborative data collation network establishment with Intranet based client-server architecture, and symptom variable trend analysis.

First, the inductive inference machine will be developed around information entropy, unbiased probability extraction using entropy maximum principle [6], and the most optimum clustering by entropy minimum principle [7]. The term "entropy" is "information entropy" which is defined as the expected value of information. The information measure compares the contents of data one receives to prior state of expectation. The higher was one's prior estimate of the probability for an outcome (or a disorder) to occur, the lower will be the information one gains by observing it to occur. Therefore, when the information quantity is the minimum, all of the information has been extracted from the available data, which leads to derive the best classification rule with the given data. Therefore, in entropy minimum state, all of the information has been extracted from the available data, which leads to derive a classification rule for affirming or negating a disorder when data samples are the only source of information [8]. Information entropy therefore is interpreted as a measure of uncertainty in classifying disorders, and entropy minimization is an ordering principle which determines which symptomatic variables used in past diagnostics of a disorder are more like those in present diagnostics with sufficiently high confidence.

A rule for diagnosis in inductive reasoning is actually a set of rules or rule patterns, each of which is formed using a single or multiple symptom variables at each step, starting from the most important variable (one with lowest entropy) for diagnosis classification. At each step, a rule pattern is generated using a single or multiple variables, and the sample data which belong to the pattern is tested to calculate and find the maximum entropy-based unbiased average probability for the rule at the step and, after that, they are removed from consideration in the next step of rule extraction. In the next step, the next best symptom variable is discovered, which has the minimum entropy value among all the symptom variables not considered from the remaining sample data as best symptom variable for disorder classification. Then the sample data which belong to the pattern of the discovered variable is tested to calculate and find the maximum entropy-based probability for the rule. After this, similarly, those data belonging to the rule pattern of the step are removed from consideration for the next step. This step goes on until all data samples are considered and all the rules are generated. One important product of

the rule extraction is the rule probability (along with the margin of error) [6] at each step of the rule generation. The rule probability and its margin of error play an essential role for each collaborator (or collaborated data) to justify the collaboration efforts in sharing their diagnosis data to improve the confidence and reliability of the global rule while minimizing the error.

Second, an Intranet-based server-client system will be developed as a test-bed for the inference machine with collected data and for the online collaborator activities, data uploading and access to the global rule extracted from the inference machine. Specifically, we will develop an Intranet from the network of the investigators' institution with a host serving computer and multiple computers as clients [9]. And database system will be installed at the server as well as a web-based data collection system. The rule extraction algorithm will be coded in to the server as the main processor of the server. Using the testbed, intermediate validation will be performed with readily available health assessment data such as Behavioral Risk Factor Surveillance System (BRFSS) data from Centers for Disease Control and Prevention [10]. The data will be divided in to several groups representing the collaborators who are limited to a certain geographical region or a certain ethnicity group or a certain culturally unique group. Then each divided sub-data will be uploaded from a distinct client machine to the test-bed to evaluate the performance of the database system, the inference machine and a rule extraction for each client as wells the global rule extraction from the aggregated data from all the clients, and the comparison of the global rule and the each local rule. Once the test-bed works with available sample data, the test-bed will be open to the Internet to include all collaborators.

Figure 1 illustrates the framework of the collaborative data collection and the automated rule extraction and update from the shared data. A collaborative database is formed in a computer server from the data entered by the collaborators. Using the entropy minimum principle, threshold values are calculated for the symptomatic variables used in the data. Then, the rule extraction is made from the database using the threshold values and the variables indexes while securing the entropy of the set of rules derived at minimum. This set of diagnosis rules with the rule reliability (i.e., probability and its margin of error), drawn with the maximum entropy principle, would be available to all collaborators. When more data are entered, the database is updated, and by the above mentioned process, a new set of rules with new rule reliability would be generated. This continuous improvement loop is the major point for improving the knowledge and understanding for a set of disorders which the collaborators aim to correctly diagnose, detect early, and prevent and timely treat.

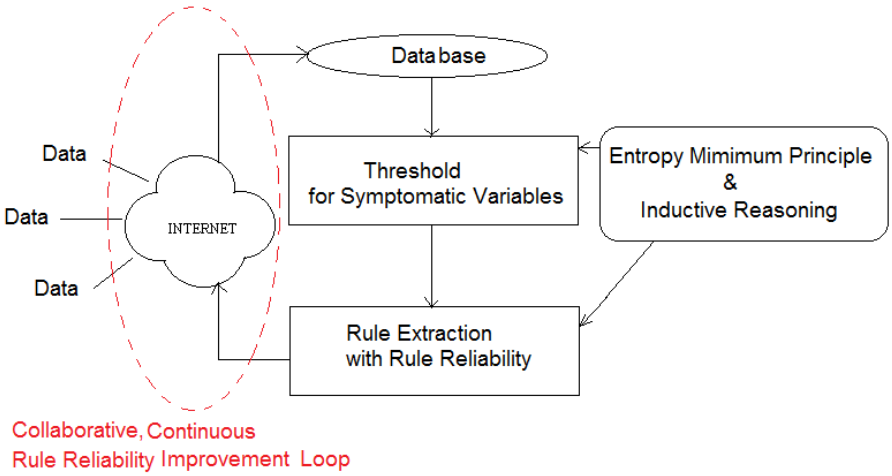

Figure 1. The Collaborative Data Collection and Automated Diagnosis Rule Generation Framework

Third, from the symptom discovery relevant to a disorder and rule extraction process, the order of importance for the symptomatic variables of higher decision confidence can be obtained from the entropy calculation. These high priority variables of lowest entropy can be used for trend analysis over time for prediction of the occurrence of the disorders or, particularly, epidemics. The planned trend analysis will be conducted using the Laplace test statistic. The Laplace trend test is a simple and powerful test for distinguishing between a constant rate at which precursory events are occurring and an increasing rate of occurrence of such event [5].

For a situation where an epidemic of a disorder is officially declared at time t_f, and m symptomatic events have been observed over a period of $[0, t_f]$ with their arrival times designated as T_1, T_2, ..., T_m, the Laplace test statistic has the following interpretation in trend analyses. Under a constant rate of symptomatic event occurrence, the arrival times to the disorder epidemic will occur randomly around the midpoint of the length of an observation interval, $t_f/2$. Therefore, the sample mean of the occurrence times will be approximately equal to $t_f/2$, hence the value of the test statistic will be small. However, if symptomatic events are occurring more frequently towards the end of the interval, the sample mean of the arrival times will be large. Therefore, in the analysis of one or a few symptomatic variables, if the Laplace test statistic is larger than the z-value of the standard normal distribution, it indicates the trend of increasing possibility of imminent onset of the disorder or epidemic. The trend analysis with shared and collaborated data would further enhance the understanding of the occurrence of the disorder of interest, and thus would help in detecting and preventing possible epidemics.

The validation includes 2 stages. The first stage of validation and analysis is done using the Intranet-based test-bed described above. Historical data of mental and behavioral assessment relevant to health/disorder status such as CDC's BRFSS will be used to test the fidelity of the entire system. A part of the data would represent a regional diagnostic data from a collaborator who is limited to a certain geographical region or a certain ethnicity group or a certain culturally unique group. Then different part of the data will be uploaded from a distinct client machine to the test-bed to verify the database system, the inference machine and its rule generation integrity for a rule extraction for each client and the global rule extraction from the aggregated data from all the clients. A comparison of the global rule and the each local rule will be made. After the test-bed validation of the system, the framework will be connected to the Internet to run a real collaborative data collection and diagnostic rule extraction system. A specific analysis for this online real-time validation includes the performance measure for the difference in the success rate of diagnosis by the global rule and that by one collaborator in a specific environment.

The most significant risk is the format incompatibility in the diverse data provided by different collaborators. The ambiguous and descriptive symptomatic variables and their values would require pre-processing for format conformity for the collaborative database on which diagnostic rule generation algorithm is based. Mitigation steps include fuzzy-logic based quantification in which descriptive terms are expressed in fuzzy membership functions and then converted to numerical values through a de-fuzzification process [11].

References

[1] American Psychological Association, "African Americans have limited access to mental and behavioral health care," http://www.apa.org/about/gr/issues/minority/access.aspx [accessed: 8/31/2017]

[2] Mild Behavioral Impairment Checklist, International Society to Advance Alzheimer's Research and Treatment (ISTAART). Available from: http://www.ctvnews.ca/polopoly_fs/1.3000545!/httpFile/file.pdf (August 2016)
[3] C. J. Kim, "An Algorithmic Approach for Fuzzy Inference," IEEE Transactions on Fuzzy Systems, Vol.5, no.4, pp.585-598, Nov 1997.
[4] C. J. Kim and B. D. Russell, "Classification of Faults and Switching Events by Inductive Reasoning and Expert System Methodology," IEEE Trans. on Power Delivery, vol. 4, no.3, July 1989. p. 1631.
[5] C. J. Kim, "Identification of Symptom Parameters for Failure Anticipation by Timed-Event Trend Analysis," Power Engineering Letter, pp.48-50, September 2000.
[6] E. T. Jaynes, Lecture Note: Probability theory with applications in Science and Engineering, February 1974. p. 16- 12 and p. 16-19.
[7] Ronald Christensen, Entropy Minimax Sourcebook Vol. II Philosophical Origin, Entropy Limited, 1980.
[8] Ronald Christensen, Entropy Minimax Sourcebook Vol III Computer Implementation. Entropy Limited, 1980.
[9] C. J. Kim and Jamshid Goshtasbi, "Internet-Based Global Decision Making Architecture", Proceedings of the ICSA 14th International Conference on Computers in Industry and Engineering, November 27-29, 2001, pp. 130-133, Las Vegas, NV
[10] Behavioral Risk Factor Surveillance System, Centers for Disease Control and Prevention, https://www.cdc.gov/brfss [Accessed 8/31/2017]
[11] C. J. Kim, "An Algorithmic Approach for Fuzzy Inference," IEEE Transactions on Fuzzy Systems, Vol.5, no.4, pp.585-598, Nov 1997.