GEMINI: An Integrative Healthcare Analytics System

Zheng Jye Ling[§], Quoc Trung Tran[†], Ju Fan[†], Gerald C.H. Koh[§],
Thi Nguyen[†], Chuen Seng Tan[§], James W. L. Yip[§], Meihui Zhang[†]
[§]National University Health System [†]National University of Singapore
[§]{zheng_jye_ling, gerald_koh, chuen_seng_tan, james_yip}@nuhs.edu.sg
[†]{tqtrung, fanj, thi, zmeihui}@comp.nus.edu.sg

ABSTRACT

Healthcare systems around the world are facing the challenge of information overload in caring for patients in an affordable, safe and high-quality manner in a system with limited healthcare resources and increasing costs. To alleviate this problem, we develop an integrative healthcare analytics system called GEMINI which allows point of care analytics for doctors where real-time usable and relevant information of their patients are required through the questions they asked about the patients they are caring for. GEMINI extracts data of each patient from various data sources and stores them as information in a *patient profile graph*. The data sources are complex and varied consisting of both structured data (such as, patients' demographic data, laboratory results and medications) and unstructured data (such as, doctors' notes). Hence, the patient profile graph provides a holistic and comprehensive information of patients' healthcare profile, from which GEMINI can infer implicit information useful for administrative and clinical purposes, and extract relevant information for performing predictive analytics. At the core, GEMINI keeps interacting with the healthcare professionals as part of a feedback loop to gather, infer, ascertain and enhance the self-learning knowledge base. We present a case study on using GEMINI to predict the risk of unplanned patient readmissions.

1. INTRODUCTION

The healthcare industry is undergoing an unprecedented information explosion [1]. At the National University Health System (NUHS), we have systematically collected a vast amount of healthcare data since 2002 and have stored it in a Computerized Clinical Data Repository (CCDR). Like other healthcare providers around the world, the increasing demand for high-quality care can be a challenge given limited healthcare resources and rising costs. Using advanced information technology (e.g., machine learning and data integration techniques) for healthcare, predictive analytics can potentially alleviate the pressure on precious resources while ensuring the quality of care that is rendered to patients. However, before we can deploy analytics in a healthcare setting, it is important to address the following problems:

Proceedings of the VLDB Endowment, Vol. 7, No. 13

1. Data of patients are stored across different systems. Hence, healthcare professionals may need to scan through several different systems to obtain relevant data. Similarly, various questions related to the monitoring of quality of care requiring routine reporting of, for example, the total number of patients readmitted into the hospital within 30 days, or the total number of diabetic patients with glycated hemoglobin (HbA1c) values more than 7%, can be a tedious and labor-intensive task, especially when data are extracted manually from various sources.

2. Many prediction tasks in the healthcare setting require prior medical knowledge, such as, identifying patients at high risk of being admitted to intensive care unit, or predicting the probability of the patients being readmitted into the hospital soon after discharge. The system needs to understand the *semantics* of the clinical data and infer implicit knowledge from the data.

To address the aforementioned problems, we develop an *integrative healthcare analytics* system called GEMINI[1] which allows point of care analytics for clinicans who need to ask questions about the patients they are caring for. The system consists of two components: PROFILING and ANALYTICS. The PROFILING component extracts data of each patient from various sources and stores them as information in a *patient profile graph*. The data sources include structured data, such as, patients' demographic data (e.g., age, gender), laboratory results (e.g., HbA1c values), and medications, and unstructured data (e.g., free-text from a doctor's note). Figure 1(a) illustrates a patient's clinical data consisting of unstructured and structured data. The patient profile graph provides a holistic and unified view of a patient's clinical data which simplifies the various routine or daily tasks performed by healthcare professionals and administrators. Figure 1(b) illustrates a profile graph constructed from the patient's clinical data in Figure 1(a). This graph contains key entities, such as, diseases (e.g., Diabetes Mellitus) and medication (e.g., Glipizide), identified from unstructured data (doctor's note) and structured data (e.g., dosage regimen of medication), and captures the relationships between these entities (e.g., Glipizide is used to treat Diabetes Mellitus). The ANALYTICS component analyzes the patient profile graphs to infer implicit information and extract relevant features for the prediction tasks. From the example of the profile graph in Figure 1(b), based on the laboratory result of HbA1c at 7.8%, we can infer that the patient's diabetes mellitus condition is not well-controlled.

[1]GEMINI stands for "**GE**neralizable **M**edical **I**nformation a**N**alysis and **I**ntegration System".

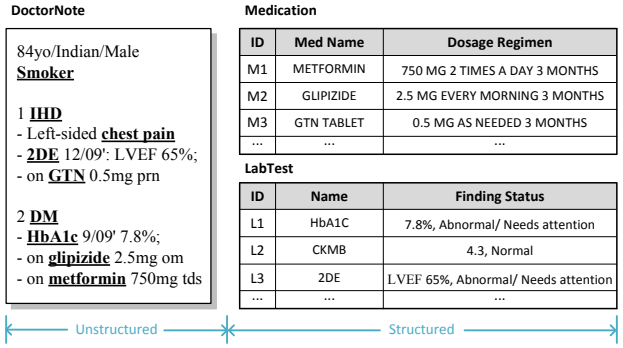
We had to address several technical challenges when developing GEMINI. First, the system needed to understand the unstructured data from doctor's note, a data source containing additional information of patient's healthcare profile [9]. There are several well-known Natural Language Processing (NLP) engines for processing clinical documents, such as, MedLEE [4] and cTAKES [10], and several medical dictionaries, such as, the Unified Medical Language System (UMLS) [2]. However, there were two issues to address:

- The text needed to be contextualized to each organization's practice, e.g., doctors in a particular department may use a different convention or notation from another department. For instance, when doctors write "PID" in the orthopaedic department, the acronym refers to "Prolapsed Intervertebral Disc" only and not "Pelvic Inflammatory Disease".
- Existing knowledge bases lack domain-specific relationships, such as, the relationship between a disease and a laboratory test. The relationships that exist between these two concepts, such as, HbA1c and Diabetes Mellitus (DM), is the use of HbA1c to monitor the control of DM, playing a crucial role in realizing the full potential of semantic computing. For instance, from the laboratory result of HbA1c, we can infer whether the DM condition is well-controlled.

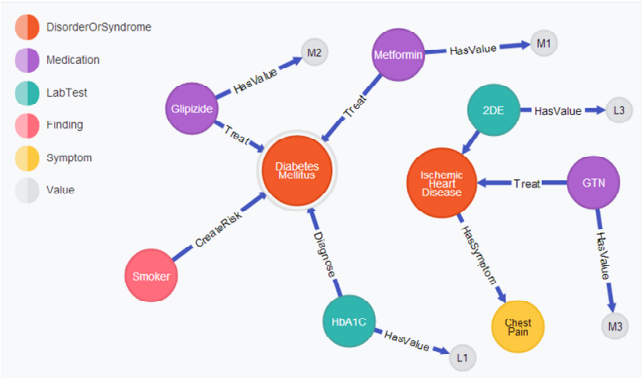
Another technical challenge is that many tasks in healthcare analytics cannot be easily solved by conventional data mining techniques. More specifically, there is usually a lack of training samples with well-defined class labels. For instance, when predicting the risk of committing suicide for each patient, the total number of patients known to have committed suicide (i.e., class 1) is very small. However, it does not mean that all the remaining patients did not commit suicide (i.e., class 0). Hence we need to infer the correct class labels for these patients.

GEMINI adopts an iterative process where the system keeps interacting with the healthcare professionals as part of a feedback loop to gather, infer, ascertain and enhance the self-learning knowledge base [8]. More specifically, to construct the patient profile graph, GEMINI leverages the information from knowledge base together with the implicit information inherent in the doctor's notes. We observe that in many notes, doctors write the related diseases, medications and laboratory tests next to each other. Hence, these patterns are captured by GEMINI as information for improving the accuracy of identifying and extracting concepts and enhancing the knowledge base. GEMINI also poses questions to the doctors for verification. Based on the answers from the doctors, GEMINI adjusts its inference results. The generation of patient profile graphs gets more accurate and complete as the system runs more iterations. Meanwhile, the knowledge base becomes more comprehensive and customized to each organization's practice. For the analytics tasks, GEMINI utilizes doctor's input to label a small number of patients with the most informative data and to provide expert rules/hypotheses, by integrating them into the analytics algorithms.

Organization. The remainder of this paper is organized as follows. Section 2 presents the system architecture of GEMINI. The next two sections discuss the two components of GEMINI: PROFILING (Section 3) and ANALYTICS (Section 4). Section 5 presents a case study of using GEMINI to predict the risk of unplanned readmissions. Finally, Section 6 concludes our work.



(a) Original Clinical Data.



(b) Patient Profile Graph.

Figure 1: Integrative Patient Profiling.

2. ARCHITECTURE OF INTEGRATIVE HEALTHCARE ANALYTICS

The architecture of GEMINI is illustrated in Figure 2. The system takes clinical data from healthcare organizations and a medical knowledge base as input, and provides integrative healthcare analytics for our target users (such as, doctors and administrators) to address their routine or daily questions.

2.1 Input and Output

Clinical Data. GEMINI uses the clinical data drawn from the CCDR of the National University Hospital. The repository has multiple sources of patient data: 1) *structured* sources containing patients' demographics, lab test results, medication history, etc., 2) *unstructured* data sources storing free-text doctor's notes. Figure 1(a) shows the clinical data of one patient, consisting of an unstructured doctor's note and records from two structured tables: `Medication` (i.e., medication history) and `LabTest` (i.e., laboratory results).

Medical Knowledge Base. GEMINI utilizes a well-known medical knowledge base UMLS [2] to interpret unstructured doctor's notes, i.e., identifying medical concepts (e.g., diabetes mellitus), and relationships between concepts (e.g., HbA1c measures control of diabetes mellitus). UMLS contains a set of concepts and a collection of relations between concepts, as shown in Figure 3(a)[2]. Specifically, each concept record consists of a concept unique identifier (CUI), a concept name, a semantic type,

[2]For simplicity, we respectively use C_i and R_i to represent concept unique identifier (CUI) and relation unique identifier (RUI), instead of using the actual identifier values in UMLS. Also, we only pick some representative properties of concepts or relations.

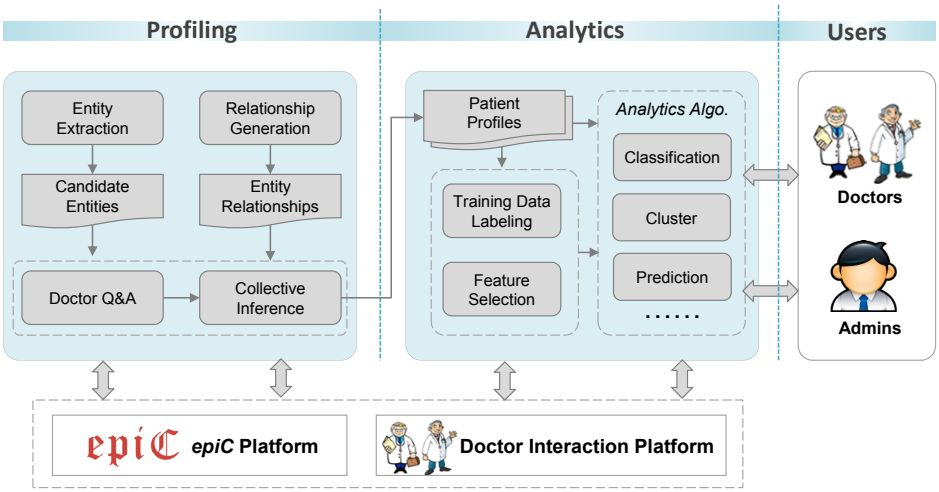


Figure 2: Integrative Healthcare Analytics System Architecture

and strings that may represent the concept. Note that a concept can be represented by multiple strings, while a string may represent multiple concepts (e.g., "DM" may refer to both concepts C_1 and C_2). Likewise, a relation record consists of a record unique identifier (RUI), the two related concepts, and the type of the relationship. For example, concept C_3 (HbA1c) has a relation named "diagnose" with concept C_1 (diabetes mellitus).

How users utilize GEMINI. Our system targets two kinds of users in healthcare organizations: (i) administrators who manage the clinical data for the daily running of the hospital, and (ii) medical professionals (e.g., doctors) who query the data for managing the clinical care of patients. It provides various analytic tasks to the users, including:

- GEMINI provides a holistic view of patient through the *patient profile graph* as shown in Figure 1(b), which contains comprehensive information of each patient. Users can interact with the graph to interrogate different facets of information on various patients. Some typical questions that a doctor might ask are: (i) list all of my patients who have hospital acquired infections; (ii) list all of my patients who are taking beta blockers treatment (where beta blockers are a class of drugs).
- GEMINI answers questions related to quality of care, such as, the total number of patients readmitted into the hospital within 30 days, or the total number of diabetic patients with glycated hemoglobin (HbA1c) values more than 7% (i.e., poor control of diabetes).
- GEMINI supports various predictive tasks, such as, identifying patients at high risk of developing heart disease in the near future, or predicting the probability that patients would re-admit into hospital within 30 days, etc.

2.2 System Components

Patient Profiling. The PROFILING component constructs a profile graph for each patient from the clinical data that provides a holistic view of the medical concepts and their relationships. For example, Figure 1 shows a profile graph and its original clinical data, and it contains not only concepts from unstructured records (i.e., the underlined words in the doctor note) and the structured records, but also various relationships between the identified concepts, such as, treat, diagnose, etc. This component utilizes NLP engines to extract named entities, called mentions. It then devises *collective inference* to simultaneously map mentions to their semantically matched concepts in the knowledge base and discovers additional relationships. To improve the accuracy of this process, the component asks doctors to verify or corroborate mention-concept mappings and concept relationships identified. In summary, the outcomes of the PROFILING component: 1) building patient profile graphs; 2) localizing and improving our medical knowledge base. Details of this profiling component are described in Section 3.

Healthcare Analytics. The ANALYTICS component provides healthcare analytics capabilities after constructing patient profile graphs. To perform the various analytic tasks of our users, such as, predicting whether the diabetic condition of a patient will be well-controlled, or whether patients will be re-admitted within 30 days, the following steps are taken by the component. It first identifies the concepts or relationships in the profile graphs that are important to the particular analytic task. This identification process can be achieved by either applying automated feature selection techniques or features selected based on the input from doctors. In addition, some analytic tasks, such as suicide prediction, may lack training data. In these scenarios, GEMINI can leverage on the expertise of doctors to label a small number of patients with the most informative data to derive a training set. The second step of the ANALYTICS component applies various analytics algorithms to the features and training data identified earlier. The analytics algorithms considered includes various classification, clustering and prediction techniques. If necessary, rules developed by experts will be incorporated to address users' analytic requirements. More details of healthcare analytics are described in Section 4.

Supporting Platforms. Two platforms are employed to support the aforementioned PROFILING and ANALYTICS components. Firstly, the clinical data in healthcare domain keeps growing dramatically. For instance, patients in intensive care unit are constantly being monitored, which would easily result in millions of records of the patents. To address the scalability issue, we utilize EPIC [6], a flexible parallel processing framework, to support:

- *distributed data storage* that effectively partitions clinical data and stores them in multiple nodes.
- *scalable NLP processing and data analytics* that involve various computation models, such as MapReduce model for

entity extraction, Pregel model for graphical inference, deep learning for analytics, etc.

The second platform is for the interaction with our domain experts, i.e., the doctors. The platform is used to publish questions to doctors and collect their expertise suggestions. For instance, as mentioned above, GEMINI can utilize the platform to leverage doctors to verify or corroborate mention-concept mappings and concept relationships. Other examples include asking doctors to label training data and identifying key features for specific analytics tasks.

3. COLLECTIVE PATIENT PROFILING

Patient profile graph. The patient profile graph is constructed from both structured and unstructured data sources and providing a holistic view of the patient profile. The graph consists of two types of nodes, namely *concept node* and *value node*. The concept nodes are represented as colored circles in the patient profile graph (see Figure 1(b)), and are constructed from the entities mentioned in the doctor notes, e.g., the concept Diabetes Mellitus is derived from the acronym DM in the free text. Each concept node is associated with a type (such as Disorder and Symptom) that is specified in UMLS. As seen in Figure 1(b), we use color to differentiate nodes based on the list of types in UMLS. The value nodes are extracted from the structured data and are mainly to attach the lab tests (e.g., HbA1c value) and the medication (dosages of medicines that the patient has ever taken). Finally, the edges in the graph represent the relationship between the connecting concept nodes. Some examples of the relationships captured in our system are listed in Table 1.

Relationship	Node1 type	Node2 type
Treat	Medication / Procedure	Disorder or Syndrome
Diagnose	Lab Test / Radiology	Disorder or Syndrome
HasSymptom	Symptom / Sign	Disorder or Syndrome
CreateRisk	Finding	Disorder or Syndrome
HasValue	Lab Test / Medication	Value

Table 1: Examples of relationships in GEMINI system.

Patient profile graph construction. To construct the patient profile graph, one option is to use NLP tools (such as cTAKES [10]) with the UMLS dictionary to extract the mentions from the doctor notes and to map the identified mentions to the UMLS concepts (i.e., the construction of concept nodes), and then create edges between the concept nodes that have the relationships specified in UMLS. However, this approach has the following limitations:

- *Ambiguous mappings*: Recall that in UMLS, one concept may correspond to multiple strings (synonyms). Similarly, one string can refer to multiple concepts, even when exact matching (instead of string fuzzy matching) is used to map the extracted mention to UMLS concepts. One example is shown in Figure 3. DM is mapped to two concepts $C1$ and $C2$ since they contain DM in the strings column. Based on our experience with the clinical data, the aforementioned approach indeed introduces a lot of spurious mappings (e.g., the red edge in Figure 3(b)).

- *Missing mappings*: The synonyms captured in the existing knowledge bases are not complete. This is because the terms used in the doctor notes could be specific within a country or a particular hospital only, whereas the existing knowledge bases might only cover the universal ones. One example that we encountered is the acronym HL in the NUH doctor note, which refers to Hyperlipidemia but not captured in UMLS. This will therefore result in the missing mappings between mentions and concepts (e.g., the grey line in Figure 3(b)).

- *Missing relationships*: We understand that the relationships between concepts covered in existing medical knowledge bases are far from complete. There are quite a number of important relationships that are missing, including the relationships between certain types of concepts (e.g., the `CreateRisk` contains relations between findings and diseases), and relationships between particular pairs of concepts (e.g., the `treat` relation between Metformin and Diabetes Mellitus denoted by the dotted line in Figure 3(b)).

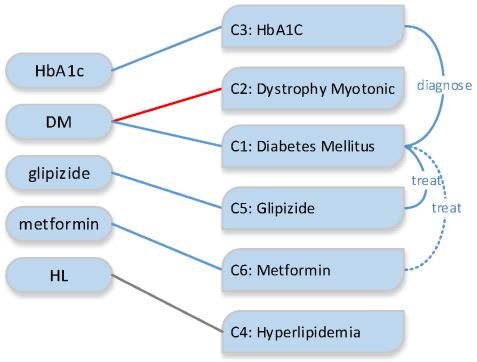
Concept

CUI	Name	Type	Strings
C1	Diabetes Mellitus	Disease or Syndrome	diabetes mellitus, DM, diabetes, ...
C2	Dystrophy Myotonic	Disease or Syndrome	Dystrophy Myotonic, DM, ...
C3	HbA1C	Laboratory Procedure	Hemoglobin A1C, HbA1C, ...
...	...	...	...

Relation

RUI	CUI1	CUI2	REL
R1	C3	C1	diagnose
R2	C5	C1	treat
R3	C7	C1	diagnose
...	...	...	...

(a) UMLS dictionary.



(b) Patient Profile Inference Illustration.

Figure 3: Collective inference for patient profiling.

Therefore, in order to build an accurate and complete patient profile graph, we need to address two essential technical challenges: 1) identifying correct mappings between mentions and concepts; 2) filling in missing relationships between the concepts.

Collective inference for patient profiling. Our main insight is that *we can improve the accuracy and the completeness of the patient profiling using the information from knowledge base together with the implicit information (signals) inherent in the doctor notes*. Consider the example in Figure 3.

- With the hints that 1) HbA1c as well as glipizide are mentioned in the doctor note immediately after DM, 2) HbA1c is a laboratory test that diagnoses diabetes, 3) glipizide is a medicine that treats diabetes, we can conjecture that DM is more likely to be $C1$ (Diabetes Mellitus) rather than $C2$ (Dystrophy Myotonic).

- Similarly for HL[3], knowing that HDL and LDL are lab tests

[3]HL is not present in the sample note in Figure 1(a) due to the space constraint. HL is listed as the third item (written in the format of `3 HL`), followed by the lab test result of HDL and LDL.

for diagnosing Hyperlipidemia, together with the pattern information of HL and HDL/LDL, it is reasonable to infer HL refers to Hyperlipidemia.

- From the doctor note (refer to Figure 1(a)), we have the knowledge that glipizide and metformin both follows DM and are written in the same/similar pattern ("- on xxx $number$ mg"). Given DM is mapped to $C1$, glipizide to $C5$, metformin to $C6$ and there is a known relation between $C5$ and $C1$, we can infer a same relationship may exist between $C6$ and $C1$ with high confidence.

However, this creates an interdependency between identifying the correct mention-concept mappings and filling in missing concept relationships. On the one hand, we need the concept relationships to identify the correct mappings. On the other hand, we need correct mention-concept mappings to infer missing relationships. Furthermore, the more comprehensive and accurate the mention-concept mappings we have, the more missing (and reliable) relationships we can discover, and vice versa. We therefore choose *a holistic approach that collectively identifies the correct mention-concept mappings and discovers the missing relationships*. Specifically, we model the task using a set of interrelated random variables following the joint probability that captures the dependencies between them, represented by a probabilistic graphical model [7]. The belief propagation inference algorithm [7] is applied to combine the diverse signals and find the optimal assignment to the variables.

Formally, we define a random variable c_m to denote the matching concept of a mention m, and $r_{cc'}$ to denote the relationship between concept c and c'. Each c_m can take a value from the set $\mathcal{C}$, which is the entire set of concepts contained in the medical knowledge base. Each $r_{cc'}$ takes a value from the set $\mathcal{R} \cup \{NA\}$ where $\mathcal{R}$ is the set of all relations captured in our system and NA denotes no relation. Following the framework of probabilistic graphical models, we define the potential functions to capture the signals we discussed above.

Mention-concept mapping: One important signal in mapping a mention to a concept is the containment of the mention string in the concept's synonym list. We define this in the following potential function:

$$\begin{aligned}\psi_{mc}(m, c_m) &= 1, \text{ if } contain(c_m, m) \text{ is true;} \\ &= 0, \text{ otherwise.}\end{aligned}$$

where $contain(c_m, m)$ is a binary feature function.[4] This potential function indicates that we prefer to mapping a mention to a concept that contains the mention. Meanwhile, we are not penalized if a mention is mapped to the other concepts, because we set the value to be 0 instead of negative. This allows the possibilities of adding missing mappings.

Concept-concept relationships: The direct signal in creating the relationships between concepts is the existence of the relationships in the knowledge base, as defined below, where $exist(c, c', r_{cc'})$ is a binary indicator of the existence of $r_{cc'}$. Again, we set the potential to be 0 for creating relationships that are potentially absent in the knowledge base, in order to allow the algorithm to discover new relationships .

$$\begin{aligned}\psi_{cc}(c, c', r_{cc'}) &= 1, \text{ if } exist(c, c', r_{cc'}) \text{ is true;} \\ &= 0, \text{ otherwise.}\end{aligned}$$

[4]One can also define the potential using normalized string similarity metrics.

Compatibility: Lastly, we define the compatibility of the assignments to the variables. Formally, we define the potential as follows for the case where m is mapped to c_m, m' is mapped to $c_{m'}$, and a relation r is created between c_m and $c_{m'}$:

$$\begin{aligned}\psi_{comp}(m, m', c_m, c_{m'}, r) &= 0, \text{ if } r \text{ is NA;} \\ &= 1, \text{ if } m, m' \text{ matches } pat(r); \\ &= -1, \text{ otherwise.}\end{aligned}$$

where $pat(r)$ returns the string patterns registered in our system for the relation type of r. For example, if r is of type treat (the relation between medication and disease), $pat(r)$ may return "m' - on m $number$ mg" as one of its possible patterns. Simpler patterns include "offset of m and m' in $range(x, y)$".[5] As we can see from the potential function, we favor the assignments that can increase the compatibility.

Collective objective: Overall, the goal is to find the assignment to the variables c_m and $r_{cc'}$ such that the following objective function is maximized:

$$\sum_m \psi_{mc}(m, c_m) + \sum_{c,c'} \psi_{cc}(c, c', r_{cc'}) + \sum_{m,m'} \psi_{comp}(m, m', c_m, c_{m'}, r_{c_m c_{m'}})$$

4. HEALTHCARE ANALYTICS

Similar to conventional data mining tools, the ANALYTICS component of GEMINI consists of three major steps: feature selection, training data labelling, and analytics algorithms.

Feature selection. Essentially, all features that are contained in the patient profile graphs can be used as features for the analytics tasks. In addition, ANALYTICS can derive implicit and also important features with expert input from the healthcare professionals. For example, doctors might suggest a specific feature in determining whether a disease is well-controlled that is very important when compared to other available features. ANALYTICS will then verify such hypotheses and revert back to the doctors with empirical evidence to support or reject their hypotheses. Such interaction is clearly beneficial to both the system and the doctors.

Training data labelling. In some prediction tasks, there is a lack of training samples with well-defined class labels. For instance, when predicting the risk of committing suicide for each patient, the total number of patients known to have committed suicide (i.e., class 1) is very small. However, it does not mean that all the remaining patients did not commit suicide (i.e., class 0). Hence we need to infer the correct class labels for these patients. To solve this issue, ANALYTICS can leverage on doctors' input to label a small number of patients with the most informative data (i.e., patient profile graphs) to derive a training set. There are two important issues here. First, doctors have different levels of confidence when answering different questions (i.e. doctors are reluctant to assess patient cases which they are not specialized in). Second, since there is so much information about patients, selecting relevant symptoms of each patient to present to the doctors in order not to overwhelm them is also a major issue.

In essence, what we need is a diverse set of labeled patients that somehow covers the whole data space as much as possible. For this purpose, ANALYTICS groups similar patient cases together

[5]The patterns can either come from the doctors' input, or from the pattern learning algorithm that is running as part of our system.

and shows these groups to doctors. The purpose is to let the doctors freely select the groups/patients that they feel more comfortable to provide the labels. In addition, for each cluster, ANALYTICS presents only the features such that the patients in the cluster have similar values on these features. In this way, we avoid overwhelming the doctors with too much information

ANALYTICS might need to ask doctors to label patient profiles in several rounds. In particular, after obtaining the training data with annotated labels, ANALYTICS applies the machine learning algorithms (to be described shortly) and then continues to pick tuples that they have low confidence in predicting their class labels and ask the doctors for their input on the class labels of these tuples.

Analytics. Based on the derived features and training data, ANALYTICS exploits conventional analytics algorithms, such as classification, clustering and prediction to perform the various analytics tasks. In addition, the doctors might have some expert rules/heuristics for the analytics tasks. For instance, a doctor might suggest a rule stating that an elderly patient who lives alone and have had several severe diseases might be readmitted into the hospital very frequently. Such kinds of rules should be integrated into the system. There are several ways to do it, such as, using majority-voting for the outputs of different rules/classifiers or combining features being used in different classifiers. ANALYTICS currently adopts the simple strategy of the former approach.

5. CASE STUDY: PATIENT READMISSION PREDICTION

This section presents our result on using GEMINI to predict the probability of patients being readmitted into the hospital within 30 days after discharge. We refer to the task as *readmission prediction* for short. GEMINI builds the patient profiling graphs from various medical data sources, including Discharge Summary, Patient Demographics, Visit and Encounter, Lab Results and Emergency Department. We focused only on the *elderly patients* (i.e., whose age is greater than 60) admitted to the hospital in 2012. There are in total 29049 elderly patients admitted to NUH in 2012, where 5658 patients readmitted within 30 days, i.e., the proportion of patients who were readmitted (i.e. class label 1) is 0.188.

GEMINI uses the following features for the prediction tasks: patients' demographics (age, gender, and race), hospital utilization (length of stay, previous hospitalizations and emergency visits), primary diagnosis and features derived from doctor's notes including laboratory results and past medical history (diseases).

We used WEKA [5] to run a 10-fold cross-validation and the Bayesian Network classifier to construct a readmission classifier[6]. Table 2 reports the accuracy of the prediction across all the 10 validation data. The result shows that our classifier can correctly predict 2585 cases that are actually readmitted. The precision and recall are 0.388 and 0.457 respectively. The result is promising when we compared it to the result handled manually by domain experts such as physicians, case managers, and nurses [3]. The recall reported in [3] is in the range [0.149, 0.306].

We would like to emphasize that this is a preliminary data analytics finding and there are many areas for further improvement to increase the accuracy of the prediction, such as, employing additional features, such as, vital signs, procedures, medications, and social factors (e.g., who are the caregivers), and applying special classifiers for highly-imbalanced data set. We also plan to study on a larger set of data.

[6]We also used other classifiers such as decision tree, rule-based classifier, SVM, etc and observe that the Bayesian Network classifier provides the best result.

	# actual class 1	# actual class 0
#predicted class 1	2585	4070
#predicted class 0	3073	19321

Table 2: The accuracy of our classifier.

6. CONCLUSION

This paper presents GEMINI, an integrative healthcare analytics system which allows point of care analytics for clinicans who need to ask questions about the patients they are caring for. GEMINI extracts data of each patient from various data sources and stores them as information in a *patient profile graph*. The patient profile graph provides a holistic and comprehensive information of patients' healthcare profile, which GEMINI can infer implicit information useful for administrative and clinical purposes, and extract relevant features for performing predictive analytics. At the core, GEMINI keeps interacting with the healthcare professionals as part of a feedback loop to gather, infer, ascertain and enhance the self-learning knowledge base.

7. ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation, Prime Minister's Office, Singapore under Grant No. NRF-CRP8-2011-08.

8. REFERENCES

[1] Ibm big data for healthcare. http://www.ibm.com.

[2] Unified medical language system. http://www.nlm.nih.gov/research/umls/.

[3] N. Allaudeen, J. L. Schnipper, E. J. Orav, R. M. Wachter, and A. R. Vidyarthi. Inability of providers to predict unplanned readmissions. *J Gen Intern Med*, 26(7):771–776.

[4] C. Friedman, P. O. Alderson, J. H. Austin, J. J. Cimino, and S. B. Johnson. A general natural-language text processor for clinical radiology. *JAMIA*, 1(2):161–174, 1994.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[6] D. Jiang, G. Chen, B. C. Ooi, K.-L. Tan, and S. Wu. epic: an extensible and scalable system for processing big data. *PVLDB*, 7(7):541–552, 2014.

[7] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[8] B. C. Ooi, K.-L. Tan, Q. T. Tran, J. W. L. Yip, G. Chen, Z. J. Ling, T. Nguyen, A. K. H. Tung, and M. Zhang. Contextual crowd intelligence. *SIGKDD Explorations*, 2014.

[9] S. Perera, C. A. Henson, K. Thirunarayan, A. P. Sheth, and S. Nair. Semantics driven approach for knowledge acquisition from emrs. *IEEE J. Biomedical and Health Informatics*, 18(2):515–524, 2014.

[10] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. K. Schuler, and C. G. Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *JAMIA*, 17(5):507–513, 2010.