RESEARCH

Integrating Semantic Medical Entity Relations for Disease Prediction Using SNOMED-CT Terminology

Mariam Daoud[1,2], Jimmy Xiangji Huang[1*], William Melek[2] and Joseph Kurian[2]

[*]Correspondence: jhuang@yorku.ca
[1]School of Information Technology York University, 4700 Keele St., M3J 1P3 Toronto, Ontario, Canada
Full list of author information is available at the end of the article

Abstract

Background: Due to the escalating quantity of digitized medical patient records, the need for medical knowledge discovery and decision support systems increases. Clinical decision support systems are designed to assist health professionals with decision making tasks such as medical diagnosis where the primary concern is to recognize disease risk and take action at the earliest signs.

Methods: In this paper, we propose a novel context-enhanced disease prediction approach based on leveraging semantic and contextual medical entity relations. Patient signs and symptoms are first mapped to SNOMED-CT concepts, which compose a feature space for disease prediction. Our major contributions in this paper consist of expanding the feature space using semantic and contextual concept relations of SNOMED-CT. For each concept, we define a medical entity context by integrating "defining" and "qualitative" medical aspects through the use of different types of semantic and contextual relationships. These latter relate a medical entity to anatomical body parts, morphological changes, severity and episodicity aspects or associate it to sequential or causal entities and qualitative interpretations.

Results: A case study is conducted on a real medical dataset where patient records are pre-annotated with diseases. We evaluate the impact of our proposed feature space on the disease prediction performance using the Naïve Bayesian, the Bayesian Network, Decision Trees and Support Vector Machines classifiers. Results reveal that added medical entities using specific relation types increase the disease prediction performance by 39.9%, 18.04%, 3.04% and 13.79% for cardiology, endocrinology, gastroenterlogy and ear/nose/throat diseases respectively.

Conclusions: This study has proven the effectiveness of semantic relations of SNOMED-CT in improving the prediction accuracy for specific diseases. A thorough study is needed for identifying the relevance of relation types with respect to disease types. This paper also demonstrated a concise approach to integrate novel features for disease prediction.

Keywords: Disease prediction; Decision support system; Semantic feature space; medical terminology

Background

The huge amount of data generated by modern medicine has motivated the drive to develop decision support systems for improving health care applications. Clinical decision support systems are designed to assist health professionals with decision

making tasks such as medical diagnosis where the primary concern is to recognize disease risk in patients and take action at the earliest signs. We regard disease prediction as a classification problem where the main challenging tasks are how to model the feature space for representing the patient data and how to exploit those features for disease prediction. There are two main research directions for disease prediction. The first research direction falls into the use of genome data [1] and the second direction makes use of clinical factors [2] or patient data. Patient data consists of, but is not limited to: dates and results of screenings, major illnesses and surgeries, lists of medicines, medicine dosages, allergies, family illnesses, clinical laboratory results, discharge summaries, etc. Clinical diagnosis refers to the process of determining or identifying a possible disease or disorder on the basis of medical signs and patient's reported symptoms, laboratory tests and other evidences.

Features in previous work are either predefined depending on clinical factors that are commonly correlated with the disease or extracted automatically based on electronic patient records content. Predefined features include categorical patient-dependent characteristics (such as age, sex, smoking, diabetes), vital examinations and lab test results (blood pressure, cholesterol level) [3, 2], risk factors categories [4], symptom finding [5] or a combination of the above features [6]. Once features are defined, disease prediction models consist of exploiting those features in supervised classification techniques [7, 8, 5, 6, 4, 9] such as Naïve bayes and bayesian network, decision trees, KNN, Neural Networks and class association rule mining.

In [2], the prediction of pulmonary disease is based on defining features related to patient's age, sex, race, smoking history and eight comorbidity variables related with the disease. A Bayesian network model composed of the aforementioned features is used to predict chronic pulmonary disease. In [1], genome predictive features have been defined for predicting Alzheimer's disease. Features are then used in a naive Bayes model that performs well in predicting a clinical outcome from large datasets.

An automatic extraction of features from patient discharge summaries is proposed in [10]. The diseases that have to be predicted are profiled and their associated symptoms and treatments using lexical and semantic resources, namely the umls medical terminology. A supervised classification technique based on Support Vectors machine combined with rules and dictionary look-up has been exploited for disease prediction. Association rule mining and class association rule mining have been also applied for diagnosis or symptom prediction [11, 5, 6], gene expression and cell type prediction [8].

Although a wide range of supervised classification techniques have been used for disease prediction, the performance of each classifier differs depending on the nature of the used collection. A comparison between class association rule mining and decision trees for disease prediction is performed in [6]. Results show that decision trees are shown to be not as adequate for artery disease prediction as association rules. On the other hand, a comparison between decision trees and Naïve Bayesian classifier has been performed in [12]. The study has shown that the Naïve Bayesian classifier is superior to inductive learning of decision trees. Disease prediction approaches also include unsupervised clustering and collaborative filtering. Predicting a patient's disease using similar patients, called collaborative filtering approach is proposed in [13, 14]. The approach in [13] uses collaborative

filtering methods to predict each patient's greatest disease risks based on their own medical history and that of similar patients. In [14], patient clustering has been adopted for predicting the likelihood of diseases for a specific patient given the clusters.

In this paper, we address the problem of clinical disease prediction given patient-reported symptoms and medical signs where patient records lack of semantic code annotation. We propose a novel context-enhanced feature space based on extracting semantic medical entities along with their semantic contexts from a medical metathesaurus, namely SNOMED-CT[1] (Systematized Nomenclature of Medicine-Clinical Terms). To the best of our knowledge, our work is the first attempt to study the impact of contextual relationships that define a medical entity for disease prediction. We define a medical entity context by integrating defining and qualitative medical aspects through the use of different types of semantic and contextual relationships. Those latter relate a medical entity to anatomical body parts, morphological changes, severity and episodicity aspects or associate it to sequential or causal entities and qualitative interpretations. We conduct experiments to answer the following questions: (1) Does enriching a medical concept through semantic and contextual relations improve disease prediction? (2) What is the impact of specific relation types on disease prediction performance?

Methods

Our approach aims at building a novel semantic feature space from patient records dataset for disease prediction. Disease prediction is regarded as a classification problem where given patient's reported sign and symptoms, the goal is to correctly determine the associated disease among a pre-defined set of diseases. Classification is a the supervised learning task of inferring a function from labeled training data. The training data consist of a set of training examples. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. For disease prediction, the patient records are the training examples and the final diagnoses contained in the patient records represent the class labels.

The main factors that affect the classification performance are the input representation of the learned function and the learning algorithm. First, the accuracy of the classification depends strongly on how the input of the learned function is represented. Typically, the input object is transformed into a feature vector, which contains a number of features that are descriptive of the object. The number of features should not contain enough information to accurately predict the output.

Our contributions in this paper are to build a semantic and contextual feature space from SNOMED-CT medical metathesaurus by leveraging semantic and contextual medical entities. The main intuition for building a semantic and contextual feature space is to cover all medical aspects associated with a medical entity, which adds discriminate features for identifying diseases. The feature space is then exploited in a supervised learning technique for predicting the most probable disease given a set of patient reported signs and symptoms.

[1] http://www.ihtsdo.org/snomed-ct/

A Semantic Feature Space Extraction

The use of the free-text fields of electronic patient records for composing the feature space based on simple terms present many limitations: the frequent use of (possibly non-standardized) acronyms, the presence of homonyms (the same word referring to two or more meanings) and synonyms (two or more words referring to the same entity). To tackle such limitations, previous work have used medical controlled vocabularies such as ICD (International Classification of Diseases) codes and UMLS(Unified Medical Language System) for representing clinical terms and findings with medical concepts.

Given patients records, we use a biomedical text mining tool, namely Metamap [15] to extract concepts of the SNOMED-CT metathesaurus. MetaMap uses a knowledge-intensive approach based on symbolic, natural-language processing (NLP) and computational-linguistic techniques to discover Metathesaurus concepts referred to in text. Our use of Metamap is to extract semantic features for representing a patient record. Although false concepts may be extracted given a patient record text, we believe that this can be tolerated by applying a feature selection approach and/or by the learning algorithm used for classification.

A patient record contains free text clinical note structured according to the following fields: Chief complaints, History of present illness (HPI), Vital examination, Order labs, Procedures and Treatment. Each patient record is associated with a disease type which is the final diagnosis provided by the physician. An example of a patient record is given in Table 1. Given the patient records, we use only chief complaints and history of present illness information from the patient records where we exclude diagnosis and treatment fields since they contain the target disease class that has to be predicted. We extract clinical finding concepts from SNOMED-CT, which represent the result of a clinical observation, assessment or judgment, and include both normal and abnormal clinical states. Examples of clinical finding concepts are "Clear sputum", "Normal breath sounds", "Poor posture". A "concept" in SNOMED-CT is a clinical meaning identified by a unique numeric identifier (ConceptId) and described via a set of words. We extract concepts that belong to the semantic category "Disorders". This category includes 12 semantic types: Acquired Abnormality, Anatomical Abnormality, Cell or Molecular Dysfunction, Congenital Abnormality, Disease or Syndrome, Experimental Model of Disease, Finding, Injury or Poisoning, Mental or Behavioral Dysfunction, Neoplastic Process, Pathologic Function, Sign or Symptom. A patient record is associated with one final diagnosis that represents a specific disease. An example of extracted concepts given the patient record of Table 1 is presented in Table 2.

Feature Expansion Using Semantic Relations

In order to add specific concepts to the feature space, we use semantic relations of SNOMED-CT that link a concept to other concepts. More precisely, each concept in SNOMED-CT is logically defined through its relationships to other concepts[2]. SNOMED CT relationships link each concept to other concepts that have a related

[2] http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_UserGuide_Current-en-US_INT_20130131.pdf

meaning. Figure 1 illustrates the concept "Pneumonia" and its relations to other concepts.

A main challenge in predicting diseases given patient's reported signs and symptoms is to represent the patient record with discriminative features specific to the associated disease. For example, "Ulcerative colitis", "crohn's disease" and "food poisoning" have common symptoms such as abdominal cramping and diarrhea. Our contributions for tackling this issue is to integrate patient reported qualitative aspects of the symptoms, such as "acute", "intermittent", "chronic", "severe", "mild" which present specific features of diseases when they present in similar patients records. We also integrate morphological changes, severity, episodical aspects of medical entities and related sequential or causal entities and qualitative interpretations.

Defining relations such as synonyms and parents aim at unifying the representation of different variants of the same medical concept. Qualitative relations aim at enriching a clinical finding with discriminative features. Associative relations such as Co-occurrent/sequential/causal relations aim at associating patients who have similar symptom patterns to the same disease type.

Our algorithm for semantic concept expansion is presented in Algorithm 1. The input to the algorithm is the set of SNOMED-CT concepts obtained through Metamap, named S. We expand S by using a specific semantic relation type r. For each concept c_i in S, we extract the concepts linked to c_i through semantic relation r. This process results in obtaining F_r. The union of all the concept sets F_r results in obtaining F which represents the complete set of expanded concepts of S using all types of relations.

Algorithm 1 Semantic feature expansion algorithm

INPUT: S is the set extracted concepts, semantic relation r
OUTPUT: Expanded set of concepts F_r using semantic relation r, Expanded set of concepts Ft using all relations
$F_r = \emptyset$, $F = \emptyset$
for each concept c_i in S **do**
  $R_{c_i} = extractRelations(c_i)$

  if $r \in R_{c_i}$ **then**
    $c_j = getRelatedConcept(c_i, r)$
    $F_r = F_r \cup c_j$
  end if
end for
$F = \cup F_r$

In the following we present the semantic relations defined in SNOMED-CT that we used for concept expansion.

- **Synonyms**: Multiple synonyms might be associated with a concept. For the medical concept "Normal breath sounds" the corresponding synonym is "Normal respiratory sounds".
- **Parents**: The meaning represented by a concept can be general, specific or somewhere in between. Concepts with different levels of granularity are linked to one another by is-a relationships. "Normal breath sounds" is linked to "Respiratory auscultation finding" via is-a relation. For each concept, we add its direct parents to the feature space.

- **Finding site**: This attribute specifies the body site affected by a condition. For example the concept "pneumonia" in Figure 1 has a "finding site" relationship to the concept "lung". For the symptom "Normal breath sounds" has finding site "Lower respiratory tract structure".
- **Co-occurrence/sequential/causal**: The relation type "Associated-with" asserts a co-occurrence between two concepts. The relations "After", "Due t"o, and "has Causative agent" asserts an interaction between concepts in which a clinical finding occurs after, is due or caused by another clinical finding. For example, in Figure 1 the concept "viral pneumonia" has a "causative agent" relationship to the concept "virus".
- **Definitional manifestations**: This attribute links disorders to the manifestations (observations) that define them. For example, "Seizure disorder" has "seizure" as an observation. Patient's reported signs and symptoms may be associated automatically to disease or disorders where adding their associated manifestations helps in adding discriminative features for specific diseases.
- **Morphology**: This attribute specifies the morphological changes seen at the tissue or cellular level that are characteristic features of a disease. For example, "Bone marrow hyperplasia" is has associated morphology "Hyperplasia".
- **Interpretation**: This attribute refers to the entity being evaluated or interpreted, when an evaluation, interpretation or judgement ((e.g., presence, absence, degree, normality, abnormality, etc.) is associated to the clinical finding. For example, the symptom "Decreased muscle tone" is linked to the observable entity "muscle tone" through "Interprets" relationship and to the finding value "Decreased" through "has-interpretation" relationship. Expanding the feature space with finding concepts allows associating patients who have similar symptom patterns to the same disease type.
- **Duration/Clinical course**: This attribute is used to represent both the course and onset of a disease. Many conditions with an acute (sudden) onset also have an acute (short duration) course. For example, "Acute amoebic dysentery" is associated with a "sudden onset and/or short duration".
- **Severity**: This attribute is used to categorize a clinical finding according to its severity. Possible values are mild, moderate, severe, fatal, etc. Adding the severity concepts associated with a clinical finding helps in discriminating between diseases that have similar symptoms such as food poisoning that has sudden abdominal pain and crohn's disease that has intermittent abdominal pain.
- **Episodicity**: This attribute is used to represent episodes of care provided by a physician or other care provider.

Disease Prediction

Disease prediction consists of predicting the likelihood of a disease on the basis of medical signs and patient-reported symptoms. We exploit the semantic and contextual features of patient's reported symptoms as described in the previous section for building the feature space. Hence, each medical sign or symptom is represented with a medical "concept" along with a "network" that relates it to general, associative and qualitative concepts. We conduct a case study by utilizing supervised learning

techniques, namely Naïve Bayesian, Bayesian Network, Decision Tree and Support Vector Machines (SVM). Naïve Bayesian relies on feature independence assumptions. Bayesian Network and decision trees rely on feature relationships. SVM relies on analyzing data and recognizing patterns used for classification.

- *Naïve Bayes*: A Naïve Bayesian classifier is a simple probabilistic classifier based on applying Bayes theorem with strong (naive) feature independence assumptions [16]. Naïve Bayesian classifiers have worked quite well in many complex real-world situations, especially when the features are not strongly correlated. In the present study, normal distribution was used for numeric attributes rather than kernel estimator [17].
- *Bayes Net*: Bayesian networks (BNs)classifier [18, 19] allows managing various forms of uncertainty via a probabilistic graphical model that represents random variables and conditional dependencies in the form of a directed acyclic graph. A Simple Estimator algorithm has been used for finding conditional probability tables for Bayes net. A K2 search algorithm was used to search network structure.
- *Decision trees*: The decision trees algorithm (J48) builds decision trees from a set of training data using the concept of information entropy [20]. We used Top-down decision tree/voting algorithm and 0.25 is used for the confidence factor. No Laplace method for tree smoothing.
- *Support Vector Machines (SVM)*: A non-probabilistic binary linear classifier that constructs one or more hyper planes to be used for classification. For training support vector classes, John Platt's sequential minimal optimization algorithm was used. Here multi-class problems are used using pair-wise classification. The parameters for SVM are set by default in Weka toolkit [21]. According to [22], the complexity parameter is set to 1. Epsilon for round off error is set to $1xE^{-12}$. PolyKernel is the set to be kernel. The tolerance parameter is set to 0.001.

Results and discussion

We evaluate our proposed approach by conducting a case study on real medical patient records dataset. The dataset is provided by Alpha Global IT healthcare company located in Canada. It contains 4 de-identified patient records repositories which covers a wide spectrum of diseases as presented in Table 3. Patient record repositories correspond to cardiology, endocrinology, gastoenterlogy and ear/nose/throat (ENT) totalling 5,120 patient records. Each disease type is associated with at least 50 and up to 150 patient records. For each patient record, the extracted concepts are enriched with their contexts using different types of relationships.

For the purpose of comparing different classification algorithms, we use Weka toolkit for performing the supervised classification of four basic classifiers: the Decision tree, the Naive Bayesian, the Bayesian Network and support vector Machines (SVM). The evaluation is performed using a 10-fold cross validation. The training dataset is used to train the classifier model and the testing dataset is used to test the classifier performance. The classification performance is evaluated using accuracy and F-measure. Accuracy is calculated as the number of testing records correctly

classified over the total number of testing records. F-measure combines precision and recall of the classification where Precision and recall are defined as [23].

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

where tp, fp and fn are the numbers of true positive, false positive and false negative predictions respectively for the considered class. $tp + fn$ is the total number of test examples of the considered class. F-measure is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

This is also known as the F_1 measure, because recall and precision are evenly weighted.

Overall disease prediction performance

In this experiment, we only use the extracted SNOMED-CT concepts to compose the feature space where concepts belong to "Disorders" category and refer to patient's reported signs and symptoms. Table 4 presents the overall classification performance for each dataset in terms of accuracy and F-measure using different classifiers. Results show that support vector machines (SVM) was the superior analysis classifier in average at both measures. The main reason is that it relies on recognizing patterns used for classification which separates the feature vectors into non-overlapping groups. Naïve Bayes was the second best classifier, where it assumes a feature independent model. We notice that the performance is variable across the datasets. The highest classification accuracy is obtained for gastroenterology dataset and the highest F-measure is obtained for Endocrinology dataset.

Impact of semantic relation types

In this experiment, we explore the impact of enriching the originally extracted medical concepts with semantic and contextual medical relations. We use support vector machines classifier (SVM) as a base classifier since it was the superior classifier in the previous experiment. Figures 2 to 5 and 6 to 9 respectively present the classification accuracy and F-measure using different types of medical relations on each dataset. Given a relation type r, the feature space F_r is composed of the originally extracted medical concepts and the expanded concepts as presented in Algorithm 1. In all the figures, the label "*no Relation*" refers to the use of original concepts for composing the feature space without concept expansion. The label "*all*" refers to the use of all relation types to expand the feature space. Based on the results, we conclude the following:

- On the cardiology dataset, 18 relation types have shown positive impact in terms of accuracy (Figure 2) and 4 relations have shown a positive impact

in terms of F-measure (Figure 6). The most positive relations are "interprets", "has-episodicity", "has-part" and "has-severity". Expanding the concepts with all types of medical relations (labelled "all") has performed best. The negative impact of some relations types could be due to a high relatedness in symptom descriptions between different diseases. When synonyms, "same" as or "replace" relation types are used, the overlapping features between diseases that present few common symptoms increases, which makes the disease type hard to identify. For example, the symptoms "Breathless" and "Palpitation" are common for 16 and 12 cardiology diseases respectively where the total number of diseases is 21.

- On the endocrinology dataset (Figure 3), all the concept relations types have shown positive impact in terms of accuracy except the relations "has-clinical-course" and "is-a" where they slightly decrease the accuracy. However, none of the relations improve the F-measure in Figure 7.
- On the gastroenterology dataset, three relation types, namely "associated-morphology-of", "has-clinical-course" and "replaces" have slightly increased the accuracy in Figure4. However, 17 relation types have improved the F-measure (Figure 8) by 7.24%, 5.79% obtained using respectively "replaces", "has finding site" and 4.35% for the relation "has episodicity".
- On the ENT dataset, all relation types have shown positive impact in terms of F-measure and accuracy except "associated-finding-of". The F-measure improvements in Figure 8 range from 8.33% for the relation "is interpreted by" to 14.58% for the relation "same as".

Disease prediction performance using all relation types

Figure 10 to Figure 13 present the improvements in terms of F-measure using all relations types on Cardiology, endocrinology, gastroenterology and ENT diseases. On the cardiology diseases (Figure 9), the highest F-measure improvements are achieved for Angina(39.92%) and Artil Fibrillation(14.08%) and Ephysema(8.08%). On the endocrinology diseases (Figure 10), the use of all relations types achieved 18.04%, 11.96% and 10.44% for Hypercalcemia, Pheochromocytoma and Carcinoid Syndrome respectively. On the gastroenterology diseases (Figure 11), the improvements are minor over diseases where the best one is 3.4% for Hepatic Cyst. On the ENT diseases (Figure 12), the highest improvements are achieved for Croup (11.77%), Laryngitis (11.82%) and Vocal cord nodule (13.79%).

Based on the results, we notice that the F-measure improvement is variable between diseases of the same specialty and between diseases of different specialties. The use of all relation types has been shown most effective on the cardiology diseases and less effective on gastroenterology diseases. The variation in improvement is due to the negative impact of some relation types on disease prediction and also to a high overlapping symptoms and signs between some diseases which makes the use of some relations less effective.

Discussion

Based on previous results, we assume that semantic relationships between concepts in a medical metathesaurus such as SNOMED-CT should be considered with care

for building a feature space used for disease prediction. In the following, we interpret the impact of some semantic relations that decrease the disease prediction when they are used to expand the feature space.

The use of synonyms has generally positive impact on disease prediction for all disease types. This metadata allows unifying the different variations of a medical entity in different patient records.

The use of is-a relationship has positively impacted disease prediction on ENT diseases. However it has a negative impact on the rest of datasets. We assume that this relation should be used carefully since a concept in SNOMED-CT can have more than one "is a" relationship to other concepts. For example, the concept "Cellulitis of foot" has two parents: "Cellulitis" and "Disorder of foot". Consequently, by adding the parent concepts will result in adding generic concepts which could minimize the specificity of a disease and creating false common concepts with other diseases.

The use of associative relationships are questionable for some diseases. The use of "associated with" has a negative impact on Cardiology diseases. The relation "associated with" associates a clinical finding with an organisme or substance as a causative agent or with another clinical finding with a relationship value "Due to" or "After". Adding concepts via "associated with" may add clinical finding concepts that are not necessary related to the patient, which could bias the classification process. However, via the causative agent value, it adds specific features of the disease to the feature space.

The use of "associated finding of" has a negative impact on ENT and gastroenterlogy diseases. This relation type links explicit contexts to their related clinical finding, procedures or event etc. For instance, contexts may refer to a family history of thyroid or past history of thyroid and it is not necessary present now. Adding clinical finding concepts for contexts may bias disease prediction since it result in integrating discriminative features of specific diseases that the patient does not have.

The relationship "has clinical course" did not improve disease prediction on cardiology and endocrinology diseases while it has a positive impact on gastroenterlogy and ENT diseases. This could be due to the fact that course and onset of cardiology and endocrinology diseases have less discriminative power than on the gastroenterlogy and ENT diseases. As a conclusion, we assume that contextual medical relations other than synonyms relations should be assessed for relevant with respect to a disease and then exploited in the disease prediction process.

Conclusions

This paper presents an exploratory study on the impact of semantic and contextual medical entity relations for disease prediction. Disease prediction is regarded as a problem of classification given patient's reported signs and symptoms. We use patient records repositories provided by Alpha Global IT where patient records are associated with cardiology, endocrinology, gastroenterology and Ear/nose/throat diseases. In order to solve the medical term synonyms and antonyms, we use SNOMED-CT for annotating all patient's reported signs and symptoms with medical concepts issued from SNOMED-CT. Then, each medical concept is expanded using different types of contextual relations predefined in SNOMED-CT. We exploit the defining relationships of SNOMED-CT, which are designed to define a medical concept

at different aspects such as associated morphology, episodicity, severity, etc. We conduct a comparative study on patient records datasets using existing classifiers, namely the Naïve Bayesian, the Bayesian Network, Decision Trees and Support Vector Machines (SVM). Experimental results revealed that SVM yields the best performance for all disease types. Our findings assume that specific aspects of a medical entity help in adding discriminative concepts that help in improving the disease classification. However, when there are too many overlapping symptoms, expanding the feature space using medical relations has a limited impact. In future work, we plan to group the relation types and perform a deep exploratory study on the impact of using a group of semantic and contextual relations for disease prediction. We intend to integrate a semantic feature selection method when expanding concepts via the SNOMED-CT relations and study its impact on disease prediction performance. We also plan to exploit more patient records information such as lab tests and procedures for disease prediction.

Author's contributions

This is a featuring work done by MD as a part of her Postdoctoral research in Alpha Global IT. JXH supervised the project and revised the manuscript. JXH, WM and JK contribute in the study design. All authors read and approved the final manuscript...

Acknowledgements

This research is supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada, Mathematics of Information Technology and Complex Systems (MITACS), Alpha Global IT and the IBM Shared University Research (SUR) Award.

Author details

[1]School of Information Technology York University, 4700 Keele St., M3J 1P3 Toronto, Ontario, Canada. [2]Alpha Global IT, Don Mills Road, Toronto, Ontario, Canada.

References

1. Wei Wei, G.F.C. Shyam Visweswaran: The application of naive bayes model averaging to predict alzheimer's disease from genome-wide data. J Am Med Inform Assoc **18**, 370–375 (2011)
2. BE, H., Y, D., IS, K., ST, W., MF, R.: Prediction of chronic obstructive pulmonary disease (copd) in asthma patients using electronic medical records. J Am Med Inform Assoc **18**, 370–375 (2011)
3. Kuttikrishnan, M.: A novel approach for cardiac disease prediction and classification using intelligent agents. CoRR **abs/1009.5346** (2010)
4. Pitt, E.: Application of data mining techniques in the prediction of coronary artery disease : use of anaesthesia time-series and patient risk factor data. PhD thesis, Queensland University of Technology (2009)
5. K. Srinivas, A.G. G. Raghavendra Rao: Mining association rules from large datasets towards disease prediction. In: Proceedings of the International Conference on Information and Computer Networks (ICICN' 2012)
6. Carlos, O.: Comparing association rules and decision trees for disease prediction. In: Proceedings of the International Workshop on Healthcare Information and Knowledge Management (HIKM '06), pp. 17–24. ACM, New York, NY, USA (2006)
7. Hu, Q.V., Huang, J.X., Melek, W., Kurian, C.J.: A time series based method for analyzing and predicting personalized medical data. In: Brain Informatics, pp. 288–298 (2010)
8. Icev, A., Ruiz, C., Ryder, E.F.: Distance-enhanced association rules for gene expression. In: Zaki, M.J., Wang, J.T.-L., Toivonen, H. (eds.) BIOKDD, pp. 34–40 (2003)
9. Soni, J., Ansari, U., Sharma, D., Soni, S.: Predictive data mining for medical diagnosis: An overview of heart disease prediction. IJCA **17**(8), 43–48 (2011)
10. yang: A text mining approach to the prediction of disease status from clinical discharge summaries. J Am Med Inform Assoc **16**, 596–600 (2009)
11. McCormick, T.H., Rudin, C., Madigan, D.: Bayesian hierarchical modeling for predicting medical conditions. The Annals of Applied Statistics **6**(2), 652–668 (2012)
12. Kononenko, I.: Inductive and bayesian learning in medical diagnosis. Applied Artificial Intelligence **7**(4), 317–337 (1993)
13. Davis, D.A., Chawla, N.V., Christakis, N.A., Barabási, A.-L.: Time to care: a collaborative engine for practical disease prediction. Data Min. Knowl. Discov. **20**(3), 388–415 (2010)
14. Paul, R., Md., A.S., Hoque, L.: Clustering medical data to predict the likelihood of diseases. In: Fifth IEEE International Conference on Digital Information Management, ICDIM 2010, July 5-8, 2010, Lakehead University, Thunder Bay, Canada, pp. 44–49 (2010)
15. Aronson, A.R.: Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. In: Proceedings of AMIA, Annual Symposium, pp. 17–21 (2001)
16. Lewis, D.D.: Naive (bayes) at forty: The independence assumption in information retrieval. In: Proceedings of the 10th European Conference on Machine Learning. ECML '98, pp. 4–15. Springer, London, UK (1998)

17. John GH, L.P.: Estimating continuous distributions in bayesian classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence
18. E., N.R.: Probabilistic Reasoning in Expert Systems: Theory and Algorithms. John Wiley & Sons, Inc., New York, NY, USA (1990)
19. Pazzani, M.J.: Searching for dependencies in bayesian classifiers. In: Learning from Data: AI and Statistics V, pp. 239–248. Springer, ??? (1996)
20. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**, 81–106 (1986)
21. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. **11**, 10–18 (2009)
22. Keerthi, S., Shevade, S., Bhattacharyya, C., Murthy, K.: Improvements to platt's SMO algorithm for SVM classifier design. Neural Comput **13**, 637–649 (2001)
23. Olson, D.L., Delen, D.: Advanced Data Mining Techniques, 1st edn. Springer, ??? (2008)

Figures

Figure 1 Illustration of relations in SNOMED-CT.

Figure 2 Prediction accuracy using relation-dependent feature expansion on cardiology diseases

Figure 3 Prediction accuracy using relation-dependent feature expansion on endocrinology diseases

Figure 4 Prediction accuracy using relation-dependent feature expansion on gastroenterology diseases

Figure 5 Prediction accuracy using relation-dependent feature expansion on ENT diseases

Figure 6 F-measure using relation-dependent feature expansion on cardiology diseases

Figure 7 F-measure using relation-dependent feature expansion on endocrinology diseases

Figure 8 F-measure using relation-dependent feature expansion on gastroenterology diseases

Figure 9 F-measure using relation-dependent feature expansion on ENT diseases

Figure 10 F-measure Improvement using all relations on cardiology diseases

Tables

Figure 11 F-measure Improvement using all relations on endocrinology diseases

Figure 12 F-measure Improvement using all relations on gastroenterology diseases

Figure 13 F-measure Improvement using all relations on Ear/Nose/Throat diseases

Table 1 An example of a text-based patient record representation

Chief Complaint	Chest Pain, Perspiration
HPI	Patients was Admitted with Complaint of Chest Pain and Perspiration. He was Feeling Nauseating. There is no History of Cough, fever, Joint Pain or shortness of breath.
Past History	Episodes of Such Chest Pain are Reported.
Diagnosis	Myocardial Infarction

Table 2 An example of a semantic-based patient record representation

Chief Complaint	C0008031,Chest Pain,[sosy];
	C0038990,Perspiration,[fndg];
HPI	C0277786,Complaint,[fndg];
	C0008031,Chest Pain,[sosy];
	C0038990,Perspiration,[fndg];
	C0262926,History,[fndg];
	C0010200,Cough,[sosy];
	C0015967,Fever,[fndg];
	C0003862,Joint Pain,[sosy];
	C0013404,Breathless,[sosy];
Diagnosis	C0027051,Myocardial Infarction,[dsyn];

Table 3 Electronic medical records dataset characteristics

Medical specialty	# of diseases	**Examples of diseases**
Cardiology	21	Aortic Stenosis, Emphysema, Cardiogenic Shock
Endocrinology	23	Ambiguous Genitalia, Carcinoid Syndrome, Impotence, Goitre
Gastroenterology	31	Hepatic Cyst, Hepatitis-B, food Poisoning,Gastro esphageal reflux disease
Ear/Nose/Throat	43	Cancer Larynx, Vertigo, Adenoid Hypertrophy, Laryngitis

Table 4 Overall disease prediction performance using only the extracted SNOMED-CT concepts

	NaïveBayes		BayesNet		Dtrees		SVM	
	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
Cardiology	66.77	0.54	58.74	0.54	65.11	0.58	70.11	0.64
Endocrinology	76.49	0.82	38.25	0.14	64.09	0.7	77.14	0.81
Gastroenterology	77.21	0.68	44.96	0.32	67.97	0.55	78.68	0.69
Ear/Nose/Throat	76.89	0.53	55.4	0.26	71.6	0.45	78.9	0.48
Average	73.106	0.668	46.062	0.258	65.658	0.578	**75.302**	**0.678**