University of Victoria | Graduate Studies

Notice of the Final Oral Examination
for the Degree of Doctor of Philosophy

of

ELHAM SEDGHI

MSc (University of Victoria, 2012)

"A Novel Stroke Prediction Model Based on Clinical Natural Language Processing"

Department of Computer Science

Monday, January 9, 2017
4:00 P.M.
Engineering and Computer Science Building
Room 468

Supervisory Committee:
Dr. Jens Weber, Department of Computer Science, University of Victoria (Co-Supervisor)
Dr. Alex Thomo, Department of Computer Science, UVic (Co-Supervisor)
Dr. Alex Kuo, School of Health Information Science, UVic (Outside Member)

External Examiner:
Dr. Tony Sahama, Electrical Engineering, Computer Science, Queensland University of Technology

Chair of Oral Examination:
Dr. Mary Ellen Purkis, School of Nursing, UVic

Dr. David Capson, Dean, Faculty of Graduate Studies

Abstract

Early detection and treatment of stroke can save lives. Before any procedure is planned, the patient is traditionally subjected to a brain scan such as Magnetic Resonance Imaging (MRI) in order to make sure he/she receives a safe treatment. Before any imaging is performed, the patient is checked into Emergency Room (ER) and clinicians from the Stroke Rapid Assessment Unit (SRAU) perform an evaluation of the patient's signs and symptoms. The question we address in this thesis is: Can Data Mining (DM) algorithms be employed to reliably predict the occurrence of stroke in a patient based on the signs and symptoms gathered by the clinicians and other staff in the ER or the SRAU? A reliable DM algorithm would be very useful in helping the clinicians make a better decision whether to escalate the case or classify it as a non-life threatening mimic and not put the patient through unnecessary imaging and tests. Such an algorithm would not only make the life of patients and clinicians easier but would also enable the hospitals to cut down on their costs.

Most of the signs and symptoms gathered by clinicians in the ER or the SRAU are stored in free-text format in hospital information systems. Using techniques from Natural Language Processing (NLP), the vocabularies of interest can be extracted and classified. A big challenge in this process is that medical narratives are full of misspelled words and clinical abbreviations. It is a well known fact that the quality of data mining results crucially depends on the quality of input data. In this thesis, as a first contribution, we describe a procedure to preprocess the raw data and transform it into clean, well-structured data that can be effectively used by DM learning algorithms. Another contribution of this thesis is producing a set of carefully crafted rules to perform detection of negated meaning in free-text sentences. Using these rules, we were able to get the correct semantics of sentences and provide much more useful datasets to DM learning algorithms.

This thesis consists of three main parts. In the first part, we focus on building classifiers to reliably distinguish stroke and Transient Ischemic Attack (TIA) from mimic cases. For this, we used text extracted from the \chief complaint" and \history of patient illness"

fields available in the patients' files at the Victoria General Hospital (VGH). In collaboration with stroke specialists, we identified a well-defined set of stroke-related keywords. Next, we created practical tools to accurately assign keywords from this set to each patient. Then, we performed extensive experiments for finding the right learning algorithm to build the best classifier that provides a good balance between sensitivity, specificity, and a host of other quality indicators.

In the second part, we focus on the most important mimic case, migraine, and how to effectively distinguish it from stroke or TIA. This is a challenging problem because migraine has many signs and symptoms that are similar to those of stroke or TIA. Another challenge we address is the imbalance that our datasets have with respect to migraine. Namely the migraine cases are a minority of the overall cases. In order to alleviate this rarity problem, we propose a randomization procedure which is able to drastically improve the classifier quality.

Finally, in the third part, we provide a detailed study on datamining algorithms for extracting the most important predictors that can help to detect and prevent Posterior circulation stroke. We compared our finding with the attributes reported by the Heart and Stroke Foundation of Canada, and the features found in our study performed better in accuracy, sensitivity and ROC.