Extracting Distinctive Features of Swine (H1N1) Flu through Data Mining Clinical Documents

Heekyong Park
Department of Biomedical Engineering,
Seoul National University
Seoul, South Korea.
care01@snu.ac.kr

Jinwook Choi, MD, PhD
Department of Biomedical Engineering,
Seoul National University
Seoul, South Korea.
jinchoi@snu.ac.kr

Abstract

Early recognition of distinguishing patterns of a novel pandemic disease is important. We introduce a methodological approach based on popular data mining techniques to extract key features and temporal patterns of swine (h1n1) flu that is discriminated from swine flu like symptoms.

1 Introduction

Early recognition of a novel pandemic is desirable to minimize its dissemination. However, it usually spends some time in developing a diagnostic test and people tend to omit the test for various reasons including cost, which leads to late recognition. Under these circumstances, symptoms and signs in clinical documents might be valuable indicators to have a population perspective on pandemic severity.

In this paper, we propose a methodological approach to extract features of swine(h1n1) flu distinctive to swine flu like patients' one.

2 Method

2.1 Data

We randomly selected twenty clinical documents from first visit records of patients who had visited emergency room in Seoul National University Hospital (SNUH) with suspected case of swine flu. Ten of the documents are RT-PCR test positive cases, which mean that the patient is swine flu infected patient, while the ten remaining documents are RT-PCR negative cases. Each document contains a patient's symptoms, observations, recent clinical histories, clinical plans, and diagnoses in natural language. The symptoms are mostly about upper respiratory infection related ones but some of the sample documents describe about ones related to other diseases.

2.2 Hypotheses

We hypothesized two things as follows. 1) We will be able to extract distinctive symptom set between swine flu and swine-flu like patients by adopting apriori association rule mining method. 2) Although the two target groups accompany similar symptoms, we will be able to make selected symptoms more discriminative by developing impact score and considering temporal aspects, development rate.

2.3 Distinguishing Symptoms Extraction

We modeled each clinical document as one *(tag, $item_1$, ..., $item_n$)* transaction of which *tag* and *items* indicate RT-PCR result and candidate features, respectively. We set symptoms, signs, travel, and contact information as candidate features. We divided target data into three groups (Table 1) and ran apriori association rule algorithm to produce rules associated with RT-PCR(+) or RT-PCR(-) cases. Then the items appeared in association rules are collected. The items are grouped into four sets according to their original rules and training data (Table 2). The union of I_{pos} and I_{pos_co} was regarded as a distinguishing feature set, I_f. To enhance discrimination ability, we used some weights to score them as shown in Table 2 and 3.

Input data set	Description
Data 1	10 h1n1 positive transactions
Data 2	10 h1n1 negative transactions
Data 3	10 h1n1 positive transactions + 10 h1n1 negative transactions

Table 1. Input data

Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 130–131,
Uppsala, Sweden, 15 July 2010. ©2010 Association for Computational Linguistics

Input data	Association rule descendant	Unique item set in association rules	Weight
Data 1	H1n1 (+)	I_{pos}	4
Data 3	H1n1 (+)	I_{pos_co}	5
Data 2	H1n1 (-)	I_{neg}	-1
Data 3	H1n1 (-)	I_{neg_co}	-2

* Subscript pos means the items are selected from h1n1(+) association rules (e.g., pos <- fever cough dyspnea) and neg means opposite cases (e.g., neg <- myalgia chilling fever). co indicates the items are selected from rules with h1n1(+) and h1n1(-) cases mixed training data (Data 3).

Table 2. Feature sets and weights

Features	I_{pos}	I_{pos_co}	I_{neg}	I_{neg_co}	score
Dyspnea	O	O			9
Pharyngeal injection	O	O			9
Sore throat	O	O		O	7
Travel	O	O		O	7
Cough	O	O	O	O	6
Fever	O	O	O	O	6
Sputum	O	O	O	O	6
Cvat (costovertebral angle tenderness)		O			5
Rhinorrhea		O			5

Table 3. Distinctive feature set(I_f) of h1n1 cases accompanied with impact scores

2.4 Disease Development Pattern Analysis

We extracted temporal information of selected features in sample documents and modeled as interval constraints. We ran Floyd-Warshall's all-pairs-shortest path algorithm to get hidden temporal relationships between two features. We traced start time gaps of the features to compare temporal patterns of the development rate of external indicators between two data sets, Data1 and Data2.

3 Result

For input data, we produced transactions and initial temporal constraint network manually. We extracted distinctive feature set, applied our scoring method, and compared temporal aspects.

We limited support value threshold as 30% for Ipos and Ineg and 0% for Ipos_co and Ineg_co due to small data size. Eight signs and symptoms as well as travel information were extracted from 47 items as distinguishing features of swine flu (Table 3). Besides sputum, pharynx injection, cvat, and travel information, five symptoms are the ones contained in the latest Centers for Disease Control (CDC) H1N1 influenza case report form. The others are strong indicators as well. The previous version of the CDC form contains sputum in the signs and symptoms section, and Himmerick (2009) described that pharyngeal injection is a clinical sign of uncomplicated swine-origin influenza A. Travel information is another important factor to diagnose an h1n1 case in Korea and included in a special purpose h1n1 clinical document format in SNUH emergency department.

We compared start time of the features but could not find any differences. The symptoms had been developed so rapidly that temporal pattern comparison between two groups was not meaningful. All the selected symptoms were developed within three days, and moreover, the symptoms in twelve documents were occurred in one day. As our data usually describes occurrence time in day granularity and the sample size is too small, we could not compare the features in finer granularity.

4 Conclusion and Discussion

In this paper, we tried to establish a methodological approach to extract distinctive features of a novel pandemic on the early symptoms experienced by the patients. We applied popular data mining techniques to swine flu suspected cases. The results correspond to outputs of previous specialized medical domain research. We could get valuable information with extremely small amount of data. This methodological approach could be used usefully in novel infectious disease management and research to prevent spreading of the pandemic at the very beginning stage.

Acknowledgement

This work was supported by the Korea Science and Engineering Foundation (**KOSEF)** grant funded by the Korea government(**MEST**) (No. 2009-0090853).

References

Kristine A. Himmerick. *H1N1 in perspective: The clinical impact of a novel influenza A virus.* JAAPA CME articles. December 01, 2009