Disease outbreak detection through clique covering on a weighted ICPC-coded graph

Klaske VAN VUURDEN, Gunnar HARTVIGSEN, Johan Gustav BELLIKA
Department of Computer Science, University of Tromsø, Norway

Abstract. Even after a decade of increased research into the problem of detecting disease outbreaks, we lack a system that can limit the number of patients affected by a potential epidemic by recognising its existence at an early stage. In this paper we suggest the use of a weighted graph representing symptoms with an exceptionally high prevalence. Cliques with high weighted edges in such a graph will represent groups of symptoms that occur together more often than usual. As a result each clique will represent the main symptoms of a disease with a high incidence rate. This will make it easier to diagnose the nature of an outbreak, to reach the affected patients at an early stage and to distinguish between outbreaks occurring simultaneously.

Keywords. Epidemiological research, population surveillance, decision support, algorithms, syndromic surveillance

Introduction

In the last decade, many population surveillance systems have been developed with the intention of detecting disease outbreaks [1,2,3]. These systems have in common that they analyse data from a number of people and use statistical methods to make a prediction about the health of a larger population. This paper argues that in order to make a precise prediction about the nature of an outbreak, the accumulated clinical symptoms of a population should be analysed as if they were experienced by a single person. Furthermore, we describe how this principle can be implemented using a weighted undirected graph. If there is sufficient information available about the symptoms of each single patient, it becomes possible to see how different symptoms relate to each other, which in turn can be used to diagnose the population. An identification of an outbreak should then be directly translated into actions towards the current patients and the potential victims of the disease.

Currently, Tromsø Telemedicine Laboratory (TTL) is working on the development of the Snow-system [4,5]. This system will process epidemiological Electronic Health Record (EHR)-data from general practitioners within a region, in order to share knowledge about prevalent contagious diseases. The Snow-system extracts, among others, data in the form of International Classification of Primary Care (ICPC)-codes.
These codes define a classification of symptoms and diagnoses. To diagnose the population of a region as a whole, the sum of all ICPC-data can be searched for groups of symptoms that occur more frequently. Such sets of symptoms could predict what the symptoms of a prevalent disease are. By considering the patient data of an entire region as opposed to that of one individual, there will be less distraction caused by noise. Noise in this context could occur if only a subset of a patient’s symptoms is being reported, if symptoms are caused by multiple conditions or if a patient receives a wrong diagnosis. By only looking at groups of symptoms with an exceptionally high frequency, a distinction can be made between different prevalent illnesses in a community.

The detection of groups will be achieved by modelling the distribution of symptoms as a weighted undirected graph, where the nodes represent symptoms, and the edges represent pair wise appearance of symptoms. A clique (a set of nodes of which each pair shares a direct relation) with a relatively high weight in such a graph, represents symptoms which are not only prevalent by themselves, but also as a group. When comparing such graphs for different periods in time, incidence rates of different diseases or phases within one disease can be explored. In practice, when cliques with an exceptionally high weight are detected, the community does not necessarily have to be alerted. Any health care professional entering new sets of symptoms into an EHR, can get a report or visualisation informing about the status of that particular set of symptoms. For example, if the set that is entered is a subset of a currently prevalent larger clique, the health care professional could use the other symptoms in that clique as a guideline for further interviewing the patient.

The idea of this system can be sketched by using the example of the outbreak of Giardiasis in Bergen 2004 [8]. The Giardia Lamblia-parasite responsible for this disease generally causes symptoms of diarrhoea, nausea, vomiting, abdominal pains, increased flatulence and weight loss. Although the last symptom is less common, a weighted graph of all symptoms would at least reveal higher weights on edges between the nodes representing the first five symptoms. The first suggestion found for these symptoms at Symptom Checker by WebMD [9] is indeed Giardiasis. The situation in
Bergen is illustrated in Figure 1 as based on the symptoms experienced by 200 Giardiasis-patients interviewed by Steen and Damsgaard [8]. The weights on the edges in this figure are the expected number of patients experiencing a pair of symptoms, based on the probability of each individual symptom occurring. The figure displays the situation if the only abnormality occurring was the Giardiasis outbreak. Since the outbreak started in October, it is likely that there was also an increase in the number of influenza patients. The typical symptoms of influenza do not overlap with those for Giardiasis. If we would add data for influenza patients, we could see a distinction between the clique for Giardiasis and that for influenza.

To demonstrate this method on a real situation, we conducted a test on data from the Central Data from General Practitioners (SEDA)-project [10], for which patient records were collected from 20 GP-clinics in Norway during one month in 2001.

The algorithm described here is based on an NP-hard problem, yet an estimation of the maximal cliques should be sufficient to point out the most significant developments in symptoms in a region. Since only the symptoms are analysed and communicated, patients described in the graph can stay relatively anonymous. Therefore, the information extracted from the graph can be used in many different applications.

1. Methods

The symptoms (in the ICPC-classification), as stated in an entry to an EHR, are encoded and sent to a regional server. Using the Snow-system [4], this data will be encoded with a minimal risk to the privacy of patients. At the regional server, a graph \( G = (V, E, W_E) \) is maintained representing current sets of symptoms throughout the region. A node \( v \) in the graph will represent symptom \( v \) and will have a label representing the number of patients with that symptom. An edge \((v_1, v_2)\) will represent the relation between symptoms \( v_1 \) and \( v_2 \). The weight \( w_{v_1,v_2} \in W_E \) represents the number of patients that have recently reported at least both symptoms \( v_1 \) and \( v_2 \). It is also possible, and probably desirable to define and maintain \( W_v \) for graph \( G \), representing the number of patients reporting each separate symptom. Adding this information to the maintained data, will not increase the space complexity and more importantly, it could offer interesting options for research. On the other hand, if this would be the only function calculated, the algorithm would fail to identify prevalent groups of symptoms in several cases. If there is a general rise in number of symptoms, it would be difficult to see how they are related and it becomes more difficult to make a diagnosis.

Graph \( G \) is a representation of all sets of symptoms that are interesting because they are prevalent. The cliques in graph \( G \) will give an approximation of the subsets that are common. Maximal cliques in \( G \) can correspond to disease descriptions. In a graph \( G' \) representing symptoms with an exceptionally high occurrence (measured against graphs constructed for similar periods), a maximal clique that covers a large amount of patients, can indicate that there is an outbreak of some disease. Indeed, if symptoms A and B both occur more frequently than usual, it is slightly more likely that they occur simultaneously in any individual. If these symptoms are typical for a certain disease, i.e., the symptoms are positively correlated, they will occur together more often when this disease is prevalent.

Ideally, all sets of symptoms occurring should be stored and examined. Collecting this information would lead to the creation of a hypergraph. Such a graph would give
an exact picture of how groups of symptoms occur in a society. However, the storage
and analysis of these hypergraphs and comparison between different sets is quite
complex. Comparison is necessary because not all patients will display the exact set of
symptoms typical for their main condition. The cliques in graph G will give an
approximation of the subsets that are common, with a smaller space complexity.

Finding an exact maximum clique is an NP-hard problem [11]. Since we are only
looking for approximate sets of symptoms, an algorithm that finds maximal cliques
with a high likelihood will be sufficient for the problem discussed in this paper. In the
experiment described in the next section, we have used an algorithm that finds some
local cliques in linear time. This algorithm makes an approximate covering of the graph
by removing detected maximal cliques from the graph until the weights of the graph
have reached some minimum. It will be possible that certain false or negative positives
occur. To account for this possibility, the results of the algorithm should be handled
carefully. The graph should be analysed often, e.g., every minute, and in a random
order to look at multiple possibilities. Even if an indication for an outbreak is found,
there should be no official alarm until it is confirmed by a health care professional. The
weight of a clique should be taken into account as an estimation of the severity of an
outbreak. It is theoretically possible that cliques are incorrectly identified as more
prevalent. These cases would be rare in real data, and the situations can be overcome
by some simple adjustments to the algorithm. Due to the scope of this paper, it will not
be possible to prove the correctness of this statement now.

To utilise the algorithm discussed here, we need to establish when a set of
symptoms is more prevalent than usual. In the next section, we assume that this
situation occurs whenever the value of an edge is higher than the measured average
value for that same edge. When examining a larger set of data it will become possible
to measure fluctuations more accurately and an edge should only be found more
prevalent if it lies outside a range of expected values.

2. The SEDA-project

To demonstrate the algorithm as described in the previous section, we implemented a
basic version. The algorithm will later on be implemented as part of the Snow-system.
At the time this paper was written, the amount of ICPC-data gathered for the Snow-
project was not sufficient to create a simulation. Instead, we performed a test on data
from the Central Data from General Practitioners (SEDA)-project [10]. This project
was conducted in 2001 by Statistics Norway to research the possibility of shared
statistics among general practitioners. The data collected for this research includes
entries to patient records from 20 GP-clinics in Norway, made in November of that
same year. (This is real data and a substitute for the data collected with Snow.)

In most of these records, ICPC-coding was limited to describing the diagnosis. Out
of the 30,565 records available, only 23,501 made use of the ICPC-codes and on
average only 1.13 codes were used per record. To run the algorithm with the SEDA-
data as it would be conducted for ICPC-symptoms, the categories of ICPC (A to Z)
were used, instead of all individual symptoms.

We used the first half of the data to obtain information about the normal frequency
of relations. Next, we compared each day in the last part of the data to the same
weekday in earlier weeks (with each data for one day representing data collected in the
last week). Cliques with an increased prevalence of at least 10 cases are shown in Table
1. The results in the table present a prediction about the increased amount of patients affected by conditions with symptoms within certain categories.

<table>
<thead>
<tr>
<th>Symptoms</th>
<th>Moment of occurrence</th>
<th>Increased number of patients affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(Psychological), R(Respiratory)</td>
<td>October 10 – 18, 2001</td>
<td>at least 10</td>
</tr>
<tr>
<td>R(Respiratory), S(Skin)</td>
<td>October 10 – 18, 2001</td>
<td>at least 10</td>
</tr>
<tr>
<td>T(Endocrine, metabolic and nutritional), L(Musculoskeletal), K(Cardiovascular)</td>
<td>October 18 – 28, 2001</td>
<td>at least 15</td>
</tr>
<tr>
<td>A(General), X(Female Genital)</td>
<td>October 20 – 27, 2001</td>
<td>at least 10</td>
</tr>
</tbody>
</table>

Table 1. Sets of symptom/diagnosis-categories with increased weight in the SEDA-data

3. Discussion

The results in Table 1 show that the algorithm described in this paper can detect cliques of prevalent symptoms from the accumulated symptoms of all symptoms reported by patients over a period of time. We cannot diagnose these results yet, since the symptom-groups are too general. However, when inspecting for example the clique of A and X-symptoms more carefully, we can conclude that the amount of pregnant women with general complaints was high in the period from the 20th to the 27th of October.

We need a data set with a higher average number of ICPC-codes to find more interesting results. To fully utilise the algorithm described in this paper, all symptoms described by patients should be documented. A possibility to overcome this problem in the Snow-system is described by Johansen et al. [4]. They propose to let patients enter their symptoms in a system while still in the waiting room. This will give patients the opportunity to describe their symptoms thoroughly without causing more workload for the physician. This method will likely cause an increase in the average number of symptoms per patient record.

It also needs to be noted here that one month of data is very little to extract reliable knowledge about patterns in the occurrence of different symptoms. Since the Snow-project will run for an indefinite period, it will likely provide an increased number of ICPC-symptoms. Experiments conducted on the data set collected by the Snow-system will verify whether the statements made in this paper hold for real data with real noise. The Snow-system will also provide us with an overview of actual occurrences of disease outbreak, which will make it possible to verify our findings.

Once implemented, the proposed method is non-intrusive. Only a minimal amount of data is needed (if only symptoms are used, the identity of individual patients is very difficult to trace) and data is only used to be inserted in the graph. The graph itself is not stored for each moment in time; it is only integrated in data representing patterns in the graph over different time periods. As a result, data will become more and more abstract over time. To get a complete picture of the health status of a community, the graph could be extended to describe diagnoses made or demographic information about patients. This would, however, make the data less anonymous and some relations detected could represent sensitive information.

The methods described in this paper could potentially be used for many different purposes. In the introduction, we touched upon the possibility of decision support when
adding a new set of symptoms to the regional graph. This implementation will provide physicians with statistical data about the current situation in their region, which can assist them in diagnosing. Other purposes could be automatic detection, visualisation of symptoms and diseases within a region, distinction between rare and common symptoms of a disease or identification of early symptoms caused by a disease. Another purpose could be that of integrating graphs for different regions with each other to form information about the global spread of pandemics. All of these possibilities could be further explored once a simple version is implemented as part of the Snow-system.

4. Conclusions
This paper examined the possibility of exploring a weighted graph in which the nodes represent clinical symptoms or diagnoses and where the edges represent pair-wise relations between the symptoms occurring in a certain region. Though the algorithm proposed is based on a complex problem, we are only searching after an estimation of all maximal cliques. With an algorithm that can approximate maximal cliques, there is a high probability that prevalent cliques will be found. With these cliques a prediction can be made about the nature of diseases currently prevalent and a distinction can be made between different independent diseases. This method will not only provide us with information about the occurrence of a disease outbreak, but it will also give an outline of the set of symptoms associated with the disease. The population affected by the disease can be approached directly and the process of diagnosing the actual disease can be accelerated.

5. Acknowledgments
This paper was partly funded by the Tromsø Telemedicine Laboratory (TTL) project 174934. We would like to thank all researchers at TTL involved in the development of the Snow-system.

References