An automated approach for finding spatio-temporal patterns in disease spread

Authors: Prathyush Sambaturu, Parantapa Bhattacharya, Jiangzhuo Chen, Bryan Lewis, Madhav Marathe, Srinivasan Venkataramanan, Anil Vullikanti

Abstract

Background: Agencies such as the Centers for Disease Control and Prevention (CDC) currently release incidence data (e.g., Influenza), along with descriptive summaries of simple spatio-temporal patterns and trends. However, public health researchers, government agencies, as well as the general public, are often interested in deeper patterns and insights into how the disease is spreading. Analysis by domain experts is needed for deriving such insights from incidence data.

Objective: Our goal is to develop an automated approach for finding interesting spatio-temporal patterns in the spread of a disease over a large region, such as: regions which have specific characteristics, e.g., high incidence in a particular week, those which showed a sudden change in incidence, or regions which have significantly different incidence compared to earlier seasons.

Methods: We develop techniques from the area of transactional data mining for characterizing and finding interesting spatio-temporal patterns in disease spread in an automated manner. A key part of our approach involves using the principle of minimum description length for characterizing a set of regions in terms of combinations of attributes; we use integer programming to find such descriptions. Our automated approach explores regions that have different kinds of temporal patterns, and ranks them based on their description length.

Results: We apply our methods for finding spatio-temporal patterns in the spread of seasonal Influenza in the United States (US) at the resolution of states. We find succinct descriptions for regions (sets of states) with specific characteristics, e.g., high activity level, which give better insight into such regions. Our approach also finds interesting patterns in the form of regions exhibiting significant changes in activity levels in a short time, and in terms of activity levels in the past seasons.

Conclusions: Our approach can provide new insights into the patterns and trends in disease spread in an automated manner. Our results show that the description complexity is an effective approach for characterizing sets of interest, which can be easily extended to other diseases and regions, beyond Influenza in the US. The patterns we find have a specific structure, which can be easily adapted for automated generation of narratives.

Keywords: Epidemic data analysis; Summarization; Spatio-temporal patterns; Transactional data mining.

Introduction

Large-scale spatio-temporal analysis and forecasts are becoming increasingly common for several diseases, such as, Influenza [1, 2, 3, 4] and Dengue [5]. There is a lot of public interest in analysis of spatio-temporal trends relating to how these diseases are spreading across the US. Such analysis includes statements about whether the season has officially started, a listing of regions which have differing levels of activity, the contrast between the current season and earlier seasons, and different kinds of trends. Such analyses have a broad readership, and are popular among news media, the general public, government agencies, as well as public health related organizations; this is evidenced by spatio-temporal patterns [6, 7] about the spread of Influenza from news agencies and blogs. For instance, the following New York Times report [6]: "For three weeks straight, the health departments of 49 states — all except Hawaii — have reported "widespread" flu activity".

Such patterns are typically identified manually by domain experts, who have significant expertise on specific diseases. Data for such analyses often comes from public health agencies, such as the Centers for Disease Control and Prevention (CDC) [8] and World Health Organization (WHO). Reports generated by CDC contain raw surveillance data on metrics, e.g., activity level from outpatient visits and rates of hospitalization, across states in the US. In addition, summaries of regions with specific characteristics, e.g., those which have high activity levels, are also included in the reports. Such summaries can be found in the CDC reports [8, 9]. For instance, the CDC report in [9] summarizes the states with high ILI activity for week ending on Mar 04, 2017 with the number of those states followed by explicit listing of their names.

Such descriptive listings are easy to construct from raw data, but are tedious to read and do not provide deeper insights into how the disease is spreading. In contrast, the analysis by Mashable [7] mentioned earlier is a *succinct* description of the set of states which have widespread activity, namely, all states in the contiguous U.S., except Oregon. The analysis by the New York Times [6] mentioned earlier is also a good and succinct description of the set of states which have reported widespread activity for three consecutive weeks. and presents succinct descriptions of the states with high ILI activity levels for the weeks of Mar 04, 2017 and Feb 10, 2018, which contrast with the simple listings by CDC mentioned earlier.

Field Code Changed

Field Code Changed

In addition to descriptions of the set of states with a particular activity level, sets exhibiting specific temporal patterns might also be of interest. An example is the set of states which maintained a stable high activity for three consecutive weeks, ending in the week of January 27, 2018:

The overall objective of our work is to automate the process of identifying
The overall objective of our work is to automate the process of identifying
The overall objective of our work is to automate the process of identifying

The overall objective of our work is to automate the process of identifying "interesting" spatio-temporal patterns from disease surveillance data, and generating succinct descriptions for them. We use the approach of mining patterns from transactional data for formalizing these questions, which has been successfully used in many areas, such as analysis of retail transaction data [10], biomedical data analysis [11, 12], and information retrieval [13].

These data sets can be encoded as a binary $n \times m$ matrix D, where the n rows represent transactions and the m columns represent items. The ith transaction corresponds to the ith row of D, and consists of a subset of items (which have value 1 in the corresponding entry). A general approach of summarizing the entries of the transaction-item matrix D is via clustering. Such clusters can then be used for pattern analysis. In [14], authors formulate the problem of summarizing a transactional dataset as an optimization and approach it via clustering and association analysis. Some later works such as [15, 16] use MDL principle to find the set of patterns that compress the dataset best
.

The main contribution of this paper is a novel approach for finding patterns in epidemic incidence data. Using the techniques of pattern mining in transactional data, we interpret the incidence data as a transaction-item matrix and develop an integer programming based technique for finding "succinct" descriptions for a given subset of regions. An automated method is designed for searching different sets of regions, and identifying patterns of interest by considering those which have the
most succinct descriptions. The Influenza incidence data for the US, obtained from CDC [8] is used to illustrate our methods. A brief
description of this data is presented in both the methods and the results sections.

Methods

Let U denote a set of elements of interest; we focus on regions, primarily states in the US, though our abstraction easily extends to other notions of regions, and other kinds of objects. We will use elements and states synonymously. There are different kinds of features associated with each state; examples of features that can be used for the Influenza data from CDC are:

- Location (e.g., Mid-Atlantic, Southwest)
- Activity level (e.g., high, moderate and low) in the tth week before the current one
- Geographical spread (e.g., widespread, local) in the tth week before the current one
- Whether the number of infections has crossed a threshold
- Whether the peak has been reached
- Similarity with past season.

All these features other than the first one (location) are epidemic specific and are computed by CDC using specific definitions. These features capture the spatial,

temporal, and severity aspects of the reported cases. We use CDC reports, e.g., [8] , to collect data that capture these features for the states in US.

We combine the data for multiple weeks (e.g., for the activity level in the tth week before the current one), and multiple seasons (e.g., for the similarity with past seasons) for our study. In general, these values are real numbers, e.g., the similarity with a past season can be a correlation metric. In this paper, we will focus on binary features (e.g., high/low activity level, which is already available in the data from CDC website). The non-binary setting can be mapped to a binary setting, through discretization of the weights.

The input data can be viewed as a matrix $D_{n\times m}$, with rows corresponding to the states, and columns corresponding to the features. We denote the jth feature as the jth column, with $D_{ij} = 1$ if state i has that feature. For instance, suppose the jth feature indicates a "mid-atlantic state". Then, $D_{ij} = 1$ for states i corresponding to New York (NY) and Pennsylvania (PA), but is 0 for California (CA). For activity levels, we have separate features for the past weeks. In addition, all the rows are also used as features (i.e., columns). The reason for this will be made clear in the description below. Table 1 shows part of an example of such a matrix D, and will be used as a running example to explain all our definitions and problem formulation.

We start with some definitions needed for formalizing this problem. Let $U = \{e_1, ..., e_n\}$ be the universe of elements, in our case, the set of all states. Let $D_j = \{i : D_{ij} = 1\}$ denote the set of elements having feature j. Let $S(j_1, \ldots, j_k) = D_{j_1} \cap ... \cap D_{j_k}$ denote the set of elements that have features $j_1, ..., j_k$; we will refer to such a set as a conjunctive *clause*. We will sometimes say that such a clause $S(j_1, \ldots, j_k)$ has length k, meaning that it is formed by the intersection of k features. We associate a cost $c(j_1, \ldots, j_k)$ function; the simplest would be $c(j_1, \ldots, j_k) = \alpha k$ for a constant α.

Term	Definition	Description
U	$\{e_1, \ldots, e_n\}$	Universe set
T	$T \subseteq U$	Target set
D_j	$\{i: D_{ij} = 1\}$	Set of elements having feature j.
$\boldsymbol{j}^{\ell}$	$j_1, \ldots, j_k$	List of features $j_1, \ldots, j_k$ in ℓ^{th} clause.

$S(\boldsymbol{j}^{\ell})$	$D_{j_1} \cap \ldots \cap D_{j_k}$	Set of elements that have all features in list $\boldsymbol{j}^{\ell}$

Table 12: Definitions and notations used in the paper

Given a target set $T \subseteq U$, we consider expressions of T in terms of unions and differences of such clauses, having the following form

$$T = \bigcup_{\ell=1}^{r} S\left(j_1^{\ell}, \ldots, j_{k_\ell}^{\ell}\right) - \bigcup_{\ell=r+1}^{s} S\left(j_1^{\ell}, \ldots, j_{k_\ell}^{\ell}\right),$$

with an associated cost of

$$\sum_{\ell=1}^{r} \alpha k_\ell + \sum_{\ell=r+1}^{s} \beta k_\ell \ ,$$

where and α and β are the constant parameters associated with positive and negative clauses. The clauses $S\left(j_1^{\ell}, \ldots, j_{k_\ell}^{\ell}\right)$ corresponding to $\ell = 1, \ldots, r$ will be viewed as "positive" clauses, and α is the cost for each such clause. The clauses corresponding to $\ell = r+1, \ldots, s$ are referred to as "negative" clauses, and describe elements which need to be removed from the set of positive clauses, in order to exactly cover the elements of T; the cost parameter corresponding to the negative clauses is β. For succinctness, we use $\boldsymbol{j}^{\ell} = \left(j_1^{\ell}, \ldots, j_{k_\ell}^{\ell}\right)$ to denote such a tuple, and $S(\boldsymbol{j}^{\ell}) = S\left(j_1^{\ell}, \ldots, j_{k_\ell}^{\ell}\right)$ as the corresponding clause. We use $NUM(\boldsymbol{j}^{\ell}) = k_\ell$ to denote the number of features involved in such a clause. Then, the representation for T can be written as

$$\mathrm{T} = \bigcup_{\ell=1}^{\mathrm{r}} \mathrm{S}(\mathbf{j}^{\ell}) - \bigcup_{\ell=\mathrm{r}+1}^{\mathrm{s}} \mathrm{S}(\mathbf{j}^{\ell}),$$

with an associated cost of

$$\sum_{\ell=1}^{r} \alpha \cdot \mathrm{NUM}(\boldsymbol{j}^{\ell}) + \sum_{\ell=r+1}^{s} \beta \cdot \mathrm{NUM}(\boldsymbol{j}^{\ell}).$$

Finally, we use $\mathcal{C}^{k} = \{\boldsymbol{j} = (j_1, \ldots, j_{k'}) : k' \leq k\}$ to denote the set of all such tuples of length at most k; for an element i, let $\mathcal{C}_i^{k} = \{\boldsymbol{j} = (j_1, \ldots, j_{k'}) : k' \leq k, i \in S(\boldsymbol{j})\}$ denote the set of tuples such that the corresponding clauses contain i.

~~**Example**. We continue our example from Table 1. An example of a target set T is the set of regions that have feature f_1, i.e., $T = \{e_1, e_4, e_5\}$. T can be expressed as combinations of different kinds of clauses. For instance, $T = S(7) \cup S(10) \cup S(11)$. This representation has cost 3α. Note that $D_2 = D_1 = T$, so T can also be expressed simply as $T = S(2)$; this has cost α. Finally, we can represent T the target set as unions and differences of clause as $T = S(2,5) \cup S(3,6) - S(3,4,6)$, since $S(2,5) = D_2 \cap D_5 = \{e_4, e_5\}$, $S(3,6) = D_3 \cap D_6 = \{e_1, e_6\}$, and $S(3,4,6) = D_3 \cap D_4 \cap D_6 = \{e_6\}$.~~

Problem Formulation

***MinDesc* problem.** Given a subset $T \subseteq U$ (referred to as a "target" set), and a dataset D, the $MinDesc(T, D)$ problem involves finding a set of tuples $\boldsymbol{j^1}, \ldots, \boldsymbol{j^s}$, such that

$$T = \bigcup_{\ell=1}^{r} S(\boldsymbol{j}^{\ell}) - \bigcup_{\ell=r+1}^{s} S(\boldsymbol{j}^{\ell}),$$

and the associated cost $\sum_{\ell=1}^{r} \alpha \cdot NUM(\boldsymbol{j}^{\ell}) + \sum_{\ell=r+1}^{s} \beta \cdot NUM(\boldsymbol{j}^{\ell})$ is minimized. In order to make the descriptions interpretable, we will restrict the sizes of these clauses, i.e., the number k_ℓ of columns whose intersection is allowed; here, we will focus on $k_\ell \leq 2$, though our approach extends to any k.

Our main idea for finding patterns of interest is to explore the space of all target sets, and identify those which have low cost descriptions. This is motivated by the *Minimum Description Length* (MDL) Principle, that forms the basis of many machine learning methods to find such descriptions; we refer to [17] for details on this topic. Specifically, we find a *succinct* representation of the set T of elements, in terms of combinations of different features. For instance, suppose the set of states which are currently experiencing high activity are precisely those in the East and South. These could be described in two alternative ways:
(a) by just listing all the states (e.g., VA, NC, MD, etc.) or,
(b) as just Eastern and Southern states. The latter is preferred because of its succinctness.

Note that for each element in U there is a feature in our data matrix. This is required to be able to represent any given target set T in terms of unions and differences of clauses.

Relaxed descriptions. As discussed in the example in , in some cases, the target set T does not have a small description, but we can find a set T' which is *close* to T, and has a smaller description than T. We model this as finding a representation for a subset T' such that $T' \approx T$, which is formalized as the $MinApproxDesc$ problem:

> Given a target set $T \subseteq U$, a dataset D, and constant parameters α, β, γ, the $MinApproxDesc(T, D)$ problem involves finding a set of tuples $\boldsymbol{j^1}, \ldots, \boldsymbol{j^s}$, such that

$$T' = \bigcup_{\ell=1}^{r} S(\boldsymbol{j}^{\ell}) - \bigcup_{\ell=r+1}^{s} S(\boldsymbol{j}^{\ell}),$$

Field Code Changed

$|\{i: i \in T \setminus T' \cup T' \setminus T\}| \leq \gamma|T|$, and the associated cost $\sum_{\ell=1}^{r} \alpha \cdot NUM(\boldsymbol{j}^{\ell}) + \sum_{\ell=r+1}^{s} \beta \cdot NUM(\boldsymbol{j}^{\ell})$ is minimized. We refer to α, β, γ as the parameters associated with the relaxed version.

In other words, the $MinApproxDesc$ problem finds a representation for a subset T' that is "close" to T.

Approach

Finding low cost descriptions for a target set T: solving the $MinDesc$ and $MinApproxDesc$ problems.

The $MinDesc$ and $MinApproxDesc$ problems are both NP-complete, even when $k_{\ell} = 1$, which corresponds to the *set cover* problem (we refer to [18] for discussion on this topic). Here, we use an integer programming approach described in the Appendix, which is able to scale well for the problems of interest in epidemic analysis. We use the Gurobi optimization software [19] to solve the resulting Integer program. The size of the instances encountered results in programs that can be solved very efficiently. So, we expect our method will scale to much larger datasets easily.

Overall workflow: finding "interesting" patterns by exploring different potential target sets. Our integer programming approach gives a low cost description for any given target set T. As mentioned earlier, from the MDL principle, a set T is likely to be an interesting pattern if it has a low cost representation—we refer to this as a "succinct description". Motivated by this idea, we discover interesting patterns by exploring many potential target sets, and ranking them based on their description cost, as we describe below.

- We consider clauses $S(\boldsymbol{j})$ with $NUM(\boldsymbol{j}) \leq k$, i.e., all possible clauses of length at most k, as potential target sets. Solve the $MinDesc$ and $MinApproxDesc$ problems for a target set $T = S(\boldsymbol{j})$; note that we will only consider representations from the set $\mathcal{C}^{k} - \boldsymbol{j}$ for such a clause $S(\boldsymbol{j})$.
- The potential target sets are partitioned into the following classes:
    - Specific features: these are individual columns D_j, e.g., the set of states with a "high activity level", "moderate activity level", etc. These correspond to the kinds of descriptions in CDC reports, as in [8].
    - Stable trends: these correspond to sets of features which are stable over time, e.g., high activity level for the past three weeks.
    - Temporal changes: those showing changes in features over a time period, e.g. low activity in one week and high in the next.
- Within each class above, the potential target sets are ordered based on the description cost, computed using the integer programming approach. We consider both the exact and relaxed representations, and retain the relaxed representation if its cost is much less.

Figure 1 shows an overview of our methodology. In the results section, we present some descriptions generated by our algorithm for a particular target set, followed

by a discussion on effects of the parameters. Finally, we present the utility of these methods in automatically identifying certain patterns in the data.

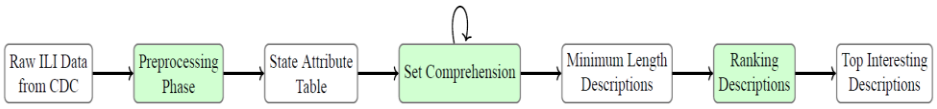

Figure 1: Overview of steps, to generate succinct description of the level of influenza like illnesses (ILI) in different states of USA. The process begins with collecting raw ILI data from CDC website, followed by the creation of a state attribute table — a domain-specific version of the transaction-item matrix D — for a given weekend. We iterate over a space of all potential target sets, and solve the MinApproxDesc problem to compute a representation. These are then ranked based on their interestingness score.

Results

Description of Dataset

We use the incidence data from CDC to prepare a dataset for any given week. In our experiments, we use data corresponding to weeks for the years 2014 to 2017. The dataset for any week has the following size:

1. Number of regions (states) or rows: 51 (50 states and District of Columbia)
2. Number of features or columns : 42 (spatial, temporal, and severity features)

Therefore, the data matrix D for a week has 2142 entries. The features can be categorized as follows:

1. **Geographical/ Spatial**: Features such as Great Lakes, South East, South West, Mid-Atlantic etc.
2. **Temporal**: Features such as Was1_high, Was2_moderate, Was52_high, etc.

We apply our $MinApproxDesc$ approach on the Influenza incidence data from the CDC. We explore the quality of descriptions, the effects of parameters, including relaxed versions, and with negative clauses. We discuss our ranking approach, and some of the trends and surprises we find. For each experiment discussed below, our algorithm ran within a few seconds.

1. Descriptions for sets of high activity levels

We use our method to compute the most succinct descriptions for the set of states with a high activity level in each week during the years 2014−2017, for different choices of the parameter γ. shows the descriptions for some of the weeks (column 2), which had succinct descriptions. The third column shows the values of γ for that description. The text description in column four is hand generated, corresponding to the solution computed using our method for the parameter value in column 3. Column 5 shows the size of the target set. While the quality of the descriptions varies quite a bit (as discussed below), this shows that our method can easily find succinct descriptions for different kinds of target sets. Using additional attributes for the regions might allow for more succinct descriptions.

Field Code Changed

2. Quality of descriptions

We attempt to evaluate the descriptions we compute in by considering the size of the target set (i.e., the set being described), and the descriptions we obtain using our methods. Some of the rows involve pretty large target sets, e.g., rows 8, 9 and 10 correspond to 13, 29 and 11 states, respectively. The CDC descriptions for these weeks would be very long lists, which are unlikely to give useful insights or identify any patterns. The description in row 9 (week 2015-01-03) is fairly succinct, and gives the following insights:

- almost all the states with high or moderate activity level in the previous week are high in the current week,
- three new states that were not experiencing high/moderate activity are now at the high activity level, and
- Florida and Georgia have experienced a sharp decline in activity levels within a week.

We note that some of the descriptions are not very insightful. For instance, the one for the week of 2016-02-20 (row 1) is simply a list. It is possible that there were no common characteristics of these states, so that the most succinct description is just a list. The description for the week of 2017-02-18 (row 5) corresponding to the parameters (0, 2, 2) is quite awkward: it combines three sets of states with different activity levels in different times in the past.

S.No.	Week	γ	Descriptions of states with high activity level in the week	Target Set Size
1	2016-02-20	0	AZ, MD, NM, TX and UT	5
2	2016-03-19	0	AR, HI, NC, NJ, VA, WY and the states with high activity both 1 week and 3 weeks ago	8
3	2016-12-24	0	AL, GA and MS	3
4	2017-01-21	0	KS, NY, WA, and states with high activity two weeks back, excluding OR and UT	10
		0.1	KS, WA, and states with high activity two weeks ago, excluding OR and UT	
		0.2	NY and states with high activity two weeks back, excluding OR and UT	
		0.3	States with high activity two weeks back excluding OR and UT	
5	2017-02-18	0	AK, IL, MD, MN, states with high activity a week ago, states with low activity two weeks ago, and	27

Field Code Changed

			states with minimal activity three weeks ago, excluding WY	
		0.3	States with high activity a week ago, excluding WY	
6	2017-03-25	0	States with high activity for last two weeks, excluding LA, MS and TX	10
7	2017-04-08	0	KY and SC	2
8	2014-12-13	0	AR, IL, IN, KS, MN, MO, OK, VA, and states with high activity a week ago	13
9	2015-01-03	0	CA, NV, NY, and states with high or moderate activity levels a week ago excluding FL and GA	29
10	2015-03-14	0	States with high activity both 1 week and 4 weeks ago, excluding CT	11

Table 23: Table describing the set of states with high activity level for certain weeks during 2014–2017. The textual description is written by hand, corresponding to the solutions computed using our method for the values shown. The abbreviations are used for state names [20]. The last column indicates the number of states with a high activity level in that week, for which the description is presented.

3. **Comparison with Pattern Mining approach:**

In this section, the utility of pattern mining on this data to find interesting patterns is explored. We consider the data corresponding to all states for a particular week as the input file. We use Weka tool to run Apriori algorithm for finding associate rules on this data. The algorithm returns several rules along with the confidence of the rule.

For instance, best rules returned by Apriori algorithm for a week:
1. low=no 27 ==> high=yes 27 <conf:(1)>
2. high=yes 27 ==> low=no 27 <conf:(1)>
3. minimal=no 27 ==> high=yes 27 <conf:(1)>
4. high=yes 27 ==> minimal=no 27 <conf:(1)>
5. moderate=no 27 ==> high=yes 27 <conf:(1)>

These rules are trivial in nature and are not very informative. A way to improve these results would be to prepare the dataset with categorical values unlike the binary (yes/no) values.

Formatted: Font: Bold

Formatted: Space Before: Auto

4. Effect of the parameters corresponding to the relaxation and cost

Recall that the parameter γ controls how accurately we attempt to describe the target set. A larger γ would mean greater error, but should lead to a more succinct description. The parameters α and β correspond to the costs of the positive and negative clauses, respectively. Setting $\beta = 0$ does not penalize negative clauses, so we expect more descriptions with negatives, compared to the case where $\beta > 0$. We examine their impact on the descriptions in .

Field Code Changed

- In row 5 corresponding to week 2017-02-18, our table shows narratives for two different sets of input parameters. We first use the regular parameters (0, 2, 2) which gives a lengthy description for the states with high activity. Next, we set $\gamma = 0.3$, and observe that the description only consists of states with high activity one week ago, excluding Wyoming. Even though it does not cover all elements of our target set, it provides information that is easier to understand. It also points out that most states that had activity during last week still continue to experience high activity levels except Wyoming.
- In row 4 corresponding to week 2017-01-21, four narratives are presented, one for each value of relaxation factor $\gamma \in \{0, 0.1, 0.2, 0.3\}$. When γ is changed to 0.1, the description omits the state New York, whereas when it is set to 0.2, it omits both Kansas and Washington. Finally, when we set $\gamma = 0.3$, it omits three states from its description. This is precisely because, the target set is of size 10 and when we set $\gamma = 0.3$, the algorithm finds a description for a target size of size at least 7. This is shown in Figure 2 where the sets in the description are represented by colors. Oregon and Utah, are not in the target set (and are excluded in the descriptions); hence, they are colored red. The figure also shows that as γ is increased, the description covers fewer states.

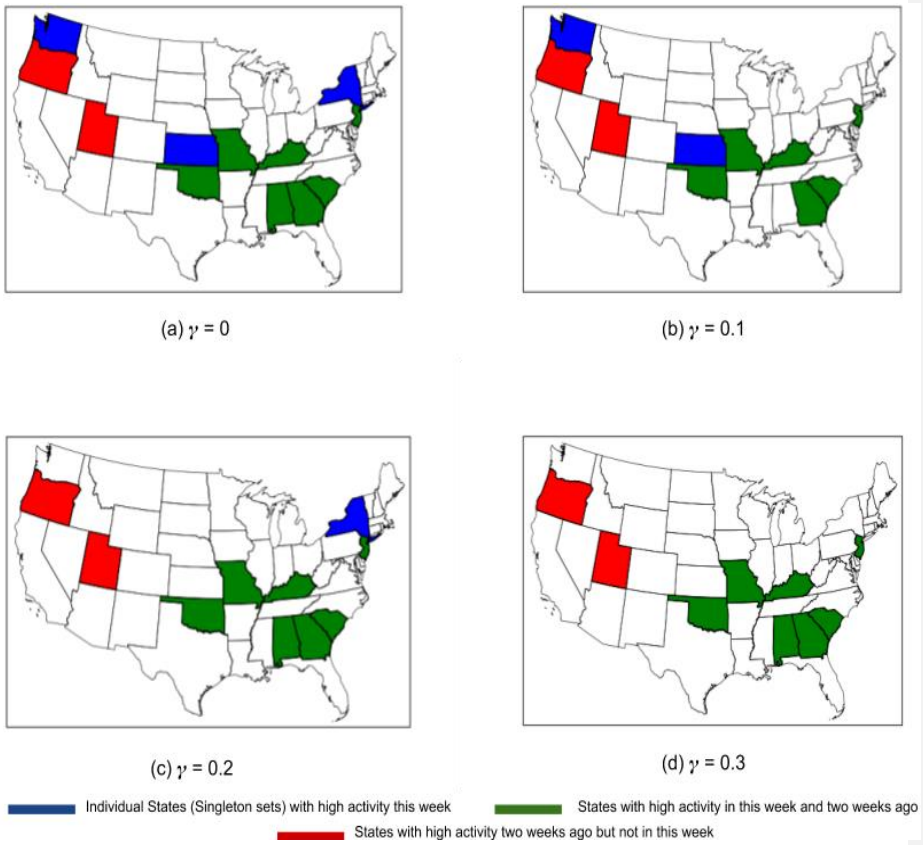

Figure 2: Effect of on the description of the set of states which high activity level in the week of 2017-01-21. The states colored green or red had high activity level two weeks ago. The states colored green or blue (New York, Kansas, Washington) in (a) are the ones with high activity level in the current week. In (b), = 0.1, and one blue colored state is dropped. Panel (c) corresponds to = 0.2, and two blue colored states are dropped and the one dropped in (b) is added. Panel (d) corresponds to 0.3, and the remaining blue state is also dropped. For = 0.3, the description only involves the green and red states.

5. Effect of negative clauses on descriptions

Some of the descriptions in exclude certain states or sets of states. This is the result of our target set representation as unions and differences of clauses. For instance, in row 5, for $\gamma = 0.3$, most states which had a high activity level a week ago, except for Wyoming, are at a high activity level currently. In row 9, the description excludes Florida and Georgia. Such negative clauses make the descriptions much simpler.

Field Code Changed

6. Generation of descriptions by ranked order

It is not known a priori which target sets would give interesting patterns. As discussed earlier, our automated procedure explores a large set of "potential" target sets corresponding to all clauses with up to k terms. We then consider a subset of these, which have descriptions of low cost, and rank them based on the potential interest for public health. We assign a score for each type of set, and construct a ranking based on the scores. Sets consisting of states with high activity level are likely to be more interesting than those with moderate, low or minimal activity levels; therefore, these are assigned scores 4, 3, 2, 1 respectively (i.e., 4 for sets with high activity level, and so on). Next, states exhibiting a sudden change in activity level (e.g., from low to high, or vice versa) are more interesting than those having no change in activity levels we assign a score of 5 for the former type, and 2 for the latter. Then, "a set of states with high this week and minimal 1 week ago" has a score of 9, while "a set of states with minimal this week and minimal 1 week ago" has a score of 3. Table 3 presents the scored descriptions for two different weeks, when run with different parameter settings. We find that the top scoring narratives generally are trends, which will be discussed next.

	Week	γ, α, β	Target set	Description	Score
1	2018-01-27	(0, 2, 2)	States with high activity this week(4), low activity two weeks ago (high to low: 5), and moderate three weeks ago (low to moderate: 5)	HI, MD, NC, OH	14
			States with moderate activity a week ago (3), minimal activity two weeks ago(moderate to minimal:5), and low three weeks ago(moderate to low: 5)	ND	13
2	2017-02-25	(0.3, 2, 4)	States with high activity this week(4), low activity a week ago (high to low: 5), moderate two weeks ago (low to moderate:5)	IA	14
			States with high activity this week (4) and low two weeks ago(high to low:5)	MD, MN	9

Table 34: Descriptions for two weeks with corresponding scores.

76. Trends

We say that a set of states has a "trend" if it exhibits a gradual increase or decrease in activity level. Examples of trend type of descriptions found by our method are:

1. *Gradual increase in the activity levels over consecutive weeks*: The states AL, GA, MS, and TN had high activity in the week of 2016-03-12, moderate the previous week, and minimal two weeks ago.
2. *Stable high activity for consecutive weeks*: In the week ending 2018-01-27, the states NJ, NM, VA, WA, WY, and the states with high activity four weeks

earlier, excluding NE and TN, had high activity levels for three consecutive weeks.

3. *Gradual decrease in ILI activity over consecutive weeks*: For the week of 2014-02-01, the activity levels in NC decreased from high to moderate to low in three consecutive weeks.

8. Surprises

We refer to a sudden rise or drop in activity levels (by at least two levels, say, from high to low, moderate to minimal, etc.) within a week's time as a surprise. Examples of surprise events identified by our methods are:

1. The activity level in NC, NM, SD, and WY jumped from low to high within a week, for the week ending 2017-02-04.
2. The activity level in NH and TN changed from high to low within a week, for the week ending 2013-02-02.

Discussion

Our results suggest that techniques from the area of transactional data mining are useful for finding spatio-temporal patterns in disease spread. In particular, the MDL principle is able to identify interesting patterns.

We find that the relaxed versions, and allowing negative clauses can significantly reduce the complexity of descriptions in many cases. Our ranking method also provides a systematic approach to identify trends and surprises in the spread of ILI. However, the descriptions of high score are not always intuitive or interesting. Instead, our ranking based approach (or other variations of it) could help provide new insights to a domain expert, who might be able to find interesting spatio-temporal patterns more easily. Thus, such an approach could be a first step in processing epidemic incidence data. We believe that including more characteristics for the data (i.e., more columns in the data matrix D) can help find more succinct descriptions. Further, the integer programming based approach is quite powerful, and more constraints can be easily added to generate descriptions with specific kinds of properties. Though the descriptions reported here were generated by hand, these are all very well structured, and could conceivably be generated using natural language processing techniques easily.

Limitations

The feature values are real numbers, e.g., the similarity with a past season can be a correlation metric, not binary. One way to handle this issue would be to map the non-binary values to binary using discretization of the weights. Since we limited our focus to only meaningful features, our current approach explores target sets with temporal properties over small time intervals. In case of an increase in number of features by a few orders of magnitude than we considered, the ILP may not be able to scale well. One way to address this problem is to design scalable heuristics that give some theoretical/ experimental guarantees.

Conclusion

Automated generation of interesting spatio-temporal patterns and trends can be very useful to public health experts, as well as the general public. Our approach, based on techniques from pattern mining, can help provide a short-list of patterns, which can then be examined more carefully by a domain expert. The techniques developed in this paper could potentially be applied for other diseases, and other public health domains.

Acknowledgements

The work of the authors has been partially supported by the following grants: NSF grant IIS-1633028, NSF grant ACI-1443054, Defense Threat Reduction Agency (DTRA) grants HDTRA1-11-1-0016 and HDTRA1-17-D-0023.

Authors' Contribution

PS, PB, BL and AV designed the study. PS, PB and AV developed the methods. All the authors helped in the evaluation and writing.

Conflicts of Interest

None

References

1. Chakraborty P, Khadivi P, Lewis B, Mahendiran A, Chen J, Butler P, Nsoesie E, Mekaru S, Brownstein J, Marathe M, Ramakrishnan N. Forecasting a moving target: Ensemble models for ILI case count predictions. SIAM International Conference on Data Mining; 2014; p. 262-270. [doi:10.1137/1.9781611973440.30]

2. Tizzoni M, Bajardi P, Poletto C, Ramasco J, Balcan D, Goncalves B, Perra N, Colizza V, Vespignani A. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. BMC Medicine; 2012: 23(1):169-214. [doi:10.1186/1741-7015-10-165]

3. Wang Z, Chakraborty P, Mekaru S, Brownstein J, Ye J, Ramakrishnan N. Dynamic poisson autoregression for influenza-like-illness case count prediction. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2015; p. 1285-1294. [doi:10.1145/2783258.2783291]

4. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible modeling of epidemics with an empirical framework; PLoS Comput Biol.; 2015. [doi:10.1371/journal.pcbi.1004382]

5. Johnson LR, Gramacy RB, Cohen J, Mordecai E, Murdock C, Rohr J, Ryan SJ, Stewart-Ibarra AM, Weikel D. Phenomenological forecasting of disease incidence using heteroskedastic gaussian processes: A dengue case study; Ann. Appl. Stat; 2018; 12(1):27–66. [doi:10.1214/17-AOAS1090]

6. "This flu season is the worst in nearly a decade, new york times, 2018. URL: https://www.nytimes.com/2018/01/26/health/flu-rates-deaths.html. [accessed 2018-11-15]. [WebCite Cache ID 73xIRUdhv] .

7. Mashable, cdc reports flu season is worsening, as 17 more children die, 2018. URL: https://mashable.com/2018/02/02/cdc-says-2018-flu-season-worse-children-deaths/#6KaneYhQEmqf. [accessed 2018-11-08]. [WebCite Cache ID 73mqMIFTH]

8. 2017-18 influenza season week 6 ending feb 10, 2018. URL: https://www.cdc.gov/flu/weekly/weeklyarchives2017-2018/Week06.htm. [accessed 2018-11-08] [WebCite Cache ID 73mqpvW7z]

9. 2016-17 influenza season week 9 ending mar 04, 2017. URL: https://www.cdc.gov/flu/weekly/weeklyarchives2016-2017/Week09.htm. [accessed 2018-11-15] [WebCite Cache ID 73xI7qtXo]

10. Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. International Conference on Very Large Data Bases (VLDB); 1994: p.487-99. [URL: http://dl.acm.org/citation.cfm?id=645920.672836]

11. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics; 2004: p.24-45. [doi: 10.1109/TCBB.2004.2]

12. Xiang Y, Jin R, Fuhry D, Dragan FF. Summarizing transactional databases with overlapped hyperrectangles. Data Min. Knowl. Discov.; 2011; 23(2): p.215–251. [doi: 10.1007/s10618-010-0203-9]

13. Wu ST, Li Y, Xu Y, Pham B, Chen P. Automatic patterntaxonomy extraction for web mining. International Conference on Web Intelligence; 2004: p.242-48. [doi: 10.1109/WI.2004.10132]

14. Chandola V, Kumar V. Summarization - compressing data into an informative representation. IEEE International Conference on Data Mining (ICDM'05); 2005. [doi: 10.1007/s10115-006-0039-1]

15. Miettinen P, Vreeken J. Model order selection for boolean matrix factorization. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD); 2011: p. 51-59. [doi: 10.1145/2020408.2020424]

16. Vreeken J, Leeuwen MV, Siebes A. Krimp: mining itemsets that compress. Data Mining and Knowledge Discovery; 2011; 23(1) p.169-214. [doi: 10.1007/s10618-010-0202-x]

17. Grünwald P. The Minimum Description Length Principle. MIT Press; 2007. URL: https://mitpress.mit.edu/books/minimum-description-length-principle. [accessed 2018-11-16]. [WebCite Cache ID 73yYgrtrL]

18. Garey MR, Johnson DS. Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman and Co.; 1979. URL: https://en.wikipedia.org/wiki/Computers_and_Intractability. [accessed 2018-11-16]. [WebCite Cache ID: 73yZKCXj2]

19. Gurobi. URL: http://www.gurobi.com/. [accessed 2018-11-08]. [WebCite Cache ID 73mAgeuFX]

20. List of us state abbreviations. URL: https://en.wikipedia.org/wiki/List_of_U.S._state_abbreviations. [accessed 2018-11-15]. [WebCite Cache ID 73xIzdXzb]

~~Appendix~~

~~The *MinDesc* problem requires exploring over the space of all possible representations for a set T, and choosing one that has the minimum cost. The *MinApproxDesc* problem has the additional requirement of ensuring that a large part of T is represented. These are both computationally very hard. Formally, these~~