A Clustering Approach in Developing Prognostic Systems of Cancer Patients

Dechang Chen  
Division of Epidemiology and Biostatistics  
Uniformed Services University of the Health Sciences  
Bethesda, MD 20814, USA  
dchen@usuhs.mil

Kai Xing  
Department of Computer Science  
The George Washington University  
Washington, DC 20052, USA  
kaix@gwu.edu

Donald Henson  
The George Washington University Cancer Institute  
The George Washington University  
Washington, DC 20037, USA  
patdeh@gwumc.edu

Li Sheng  
Department of Mathematics  
Drexel University  
Philadelphia, PA 19104, USA  
lsheng@math.drexel.edu

Arnold M. Schwartz  
Department of Pathology  
The George Washington University Medical Center  
Washington, DC 20037, USA  
aschwartz@mfa.gwu.edu

Xiuzhen Cheng  
Department of Computer Science  
The George Washington University  
Washington, DC 20052, USA  
cheng@gwu.edu

Abstract

Accurate prediction of survival rates of cancer patients is often key to stratify patients for prognosis and treatment. Survival prediction is often accomplished by the TNM system that involves only three factors: tumor extent, lymph node involvement, and metastasis. This prediction from the TNM has been limited, mainly because other potential prognostic factors are not used in the system. Based on availability of large cancer datasets, it is possible to establish powerful prediction systems by using machine learning procedures and statistical methods. In this paper, we present a clustering based approach to develop prognostic systems of cancer patients. Our method starts with grouping combinations that are formed using levels of factors recorded in the data. The dissimilarity measure between combinations is obtained through a sequence of data partitions produced by multiple clusterings. This dissimilarity measure is then used with a hierarchical clustering method in order to find clusters of combinations. Prediction of survival is made simply by using the survival function derived from each cluster. Our approach admits multiple factors and provides a practical and useful tool in outcome prediction of cancer patients. A demonstration of use of the proposed method is given for lung cancer patients.

1 Introduction

Accurate prediction of outcomes or survival rates of cancer patients is often key to stratify patients for prognosis and treatment. Outcomes of patients are usually generated using standard survival functions and various factors recorded in the database (such as SEER [10] or NCDB [9]) that have prognostic potential. All prognostic factors become integrated through the determination of outcome according to the survival rate. This integration leads to a prognostic system that can be used to predict outcome of any new patients. Clearly, a crucial question is: how can one form a powerful prognostic system for cancer patients? The traditional answer to this question is to use the TNM system [5] that involves only three factors: tumor extent, lymph node involvement, and metastasis. However, the outcome prediction from the TNM has been limited, mainly because any other potential prognostic factors are not used in the system.

In this paper, we propose a computer-based prognostic system for cancer patients that admit multiple prognostic factors. Here is idea of our approach: (i) we partition patients from a cancer dataset into “natural” groups such that patients in the same group are more similar in survival than patients from different groups; (ii) once “natural” groups are obtained, a survival function for each group can be es-
timed by a standard method. Our prognostic system then consists of groups of patients and survival functions associated with the groups.

The first step (i) is the key to the entire process. Mathematically, this step is equivalent to performing a cluster analysis on a cancer dataset. However, this type of cluster analysis is different from traditional clustering approaches, which may be elaborated below. Suppose, after some simple management, a typical record for a patient contained in a cancer dataset is of the form: \( X, X_1, \ldots, X_m \), where \( X \) is the recorded patient's survival time, which can be a censored time, and \( X_1, \ldots, X_m \) are measurements made on \( m \) risk factors or variables such as tumor size, gender, age, etc. Cluster analysis rising in (i) means that clusters of patients are sought such that patients in the same cluster are more similar in their lifetime \( T \) than patients from different groups. Here the connection between \( T \) and the observed time \( X \) is described as follows: \( T = X \) if \( X \) is an actual time to death due to the cancer under study; \( T > X \) otherwise (in this case \( X \) is a censored time.). This type of cluster analysis is not equivalent to partitioning the set of vectors \( \{ (X, X_1, \ldots, X_k) \} \) or the set \( \{ (X_1, \ldots, X_k) \} \) which could be suggested by traditional clustering methods.

The above discussed difference between the cluster analysis in (i) and the traditional clustering indicates that clustering required in (i) may not be a trivial task. Other potential challenges in accomplishing (i) include presence of a high percentage of censored observations, different types of risk factors or variables, and a large dataset size [1] [2] [11]. For example, a SEER dataset of lung cancer patients diagnosed from 1973 through 2002 has more than 500,000 patients, comprises more than 30% records with censored survival times, and involves more than 80 variables that are either on the continuous, or ordinal, or nominal scale.

To overcome the above mentioned possible difficulties, we consider subsets of a cancer data, based on combinations of levels of some known key factors. This reduces the complexity in establishing prognostic systems. We then group these subsets by a hierarchical clustering algorithm, where the distance measure between two subsets is learnt using a partition clustering method.

The rest of the paper is organized as follows. In Section 2, we present our algorithm of clustering of cancer data. An application of our algorithm to establishing a prognostic system for lung cancer patients is provided in Section 3. And finally our conclusion is given in Section 4.

2 Algorithm of Clustering of Cancer Data

A key issue related to clustering is how one measures the dissimilarity between objects. Most clustering algorithms presume a measure of dissimilarity. For example, the \( K \)-means clustering uses Euclidean distance as a dissimilarity measure. Since cancer data involve censored survival times, a direct use of existing clustering algorithms is not applicable. With cancer data, it is important to find a way to define objects and dissimilarity between objects prior to execution of any clustering algorithm.

Suppose a certain number of factors have been selected, from a data set, for consideration. Various combinations can then be formed by using levels of factors. Specifically, a combination is a subset of the data that correspond to one level of each factor. Suppose there are available a total of \( N \) combinations \( x_1, x_2, \ldots, x_n \). A combination plays a role of an object in the cluster analysis. When developing a prognostic system, we need to find groups of patients such that patients within each group are more similar in survival than patients from different groups. Assuming that all patients coming from the same combination have a similar survival rate, then the above is equivalent to finding natural groups of combinations.

After objects become available, we can start to define a dissimilarity measure between objects. A dissimilarity measure \( dis(x_i, x_j) \) is a non-negative function that is symmetric with respect to \( x_i \) and \( x_j \). For cancer data, a direct method is to define the dissimilarity between two combinations in light of the difference between the two corresponding survival functions, and the details follow below. Given two combinations \( x_i \) and \( x_j \), testing if there is a difference between the corresponding two survival functions can be done by conducting a commonly used test such as the log-rank test ([8]). It is known that the smaller the value of a test statistic, the stronger the evidence of no difference. Thus we can define dissimilarity or “distance” between \( x_i \) and \( x_j \) to be

\[
dis_0(x_i, x_j) = \text{the value of a test statistic.} \quad (1)
\]

Clearly, \( dis_0(x_i, x_j) > 0 \). This dissimilarity measure in (1) is not the one we actually use when developing cancer predictive systems. In fact, we will use the dissimilarity measure (1) for the PAM algorithm [7] only and generate a learnt dissimilarity measure through combining assignments from multiple clusterings based on the PAM. A learnt measure should be more realistic than that in (1). This learnt dissimilarity will then be used with a hierarchical clustering algorithm to produce prognostic systems.

Below we first discuss learning dissimilarity from the use of PAM. And then we present an ensemble clustering algorithm using the learnt dissimilarity and linkage methods to develop prognostic systems for cancer patients.

Learning Dissimilarity from Data

Different choices of dissimilarity functions can lead to quite different clustering results. Prior knowledge is often helpful in selecting an appropriate dissimilarity measure for
a given problem. However, it is possible to learn a dissimilarity function from the data. We describe such a procedure as follows.

Partitioning methods are usually not stable in the sense that the final results often depend on initial assignments. However, if two objects are assigned to the same cluster by a high percentage of the times of use of the same partitioning method, it is then very likely that these two objects come from a common “hidden” group. This heuristic implies that the “actual” dissimilarity between two objects may be derived by combining the various clustering results from repeated use of the same partitioning technique. Here we formalize this combining process using the PAM partitioning method.

For the data \( \{x_1, x_2, \ldots, x_n\} \), we can select \( K \) initial medoids and then run PAM with the dissimilarity measure (1) to partition the data into \( K \) clusters. It is known that the final assignment usually depends on the initial reallocation. Now we run PAM \( N \) times. Each time a number \( K \) is randomly picked from a given interval \([K_1, K_2]\). By doing this, we may end up with \( N \) possibly different final assignments. Given two objects \( x_i \) and \( x_j \), let \( p_{ij} \) denote the probability that they are not placed into the same cluster by the final assignment of a run of PAM. This probability \( p_{ij} \) can be estimated by using the results of repeated PAM clustering method. Define \( \delta_{ij} = 1 \) if the \( l \)th use of the PAM algorithm does not assign \( x_i \) and \( x_j \) into the same cluster; and \( \delta_{ij} = 0 \) otherwise. Then \( \delta_{1(i,j)}, \delta_{2(i,j)}, \ldots, \delta_{N(i,j)} \) are iid Bernoulli\( (p_{ij}) \). It is well known that the best unbiased estimator of \( p_{ij} \) is \( \sum_{l=1}^{N} \delta_{l(i,j)}/N \). This estimate will be used as the dissimilarity measure between \( x_i \) and \( x_j \), i.e.,

\[
dis(x_i, x_j) = \frac{\sum_{l=1}^{N} \delta_{l(i,j)}}{N}.
\]

A smaller value of \( \dis(x_i, x_j) \) is expected to imply a bigger chance that \( x_i \) and \( x_j \) come from the same “hidden” group.

### Clustering of Cancer Data

With the learnt dissimilarity (2) between the combinations, we can choose a clustering method to form “natural” groups of the combinations. For flexibility and easy interpretation in practice, we choose a hierarchical clustering approach. Figure 1 shows our final ensemble algorithm of clustering of cancer data (EACCD). Here the word ensemble refers to the sequence of the PAM procedures involved in the method.

Early issues on ensemble clustering were discussed in [4] from the perspective of evidence accumulation. The work in [3] combined the \( K \)-means algorithm and linkage methods to form an ensemble method of discovering sample classes using gene expression profiles.

---

**Figure 1. Ensemble algorithm of clustering of cancer data.**

1. Given \( N, K_1, \) and \( K_2 \), run the PAM clustering method \( N \) times with each \( K \) randomly chosen from \([K_1, K_2]\).
2. Construct the pairwise dissimilarity measure \( \dis(x_i, x_j) \) by using the equation (2).
3. Cluster the \( n \) objects by applying a linkage method and the dissimilarity measure \( \dis(x_i, x_j) \) from Step 2.

### 3 Results on Lung Cancer

#### Dataset

In this study, we used the SEER data [10] containing records of lung cancer patients diagnosed from the year 1988 through 1998 and examined the following factors: AJCC stage, grade, histological type, and gender. We considered four factors, \( X_1, X_2, X_3, \) and \( X_4 \) that were set to be stage, grade, histological type, and gender, respectively.

For simplicity, we only investigated the following four important levels of \( X_3 \): adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and small cell carcinoma. The levels of other three variables were those commonly used in the lung cancer study. Factor \( X_1 \) had four levels: I, II, III, and IV; factor \( X_2 \) had four levels: 1, 2, 3, and 4; and factor \( X_4 \) had two levels: 1 (male) and 2 (female). The final data we actually used involve 90, 214 patients. A portion of the data, in terms of \( X \) (survival time), \( X_1 \), \( X_2 \), \( X_3 \), and \( X_4 \), is provided in Table 1.

#### Setting of the Algorithm

To run our algorithm EACCD, we chose parameters as follows. The choice of \( N \) depends on the rate at which \( \dis \) in (2) converges to \( p_{ij} \). A large number should be chosen for \( N \), and for this purpose we set \( N = 10000 \). Any theoretically possible choices of \( K \) was used in running PAM, and thus we set \( K_1 = 2 \) and \( K_2 = 79 \), due to availability of 80 objects. In addition, the log-rank test was used to obtain the measure (1) for the PAM algorithm. And the average linkage was employed as a hierarchical clustering method.
Table 1. Lung cancer data of 90,214 patients. Survival time is measured in months. Here, adeno, squamous, large, and small represent adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and small cell carcinoma, respectively.

<table>
<thead>
<tr>
<th>Patient</th>
<th>Survival time (X)</th>
<th>Stage (X1)</th>
<th>Grade (X2)</th>
<th>Histology (X3)</th>
<th>Gender (X4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>64</td>
<td>1</td>
<td>2</td>
<td>squamous</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>1</td>
<td>3</td>
<td>large</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>2</td>
<td>3</td>
<td>squamous</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>squamous</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>3</td>
<td>3</td>
<td>squamous</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>90214</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>squamous</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2. A list of 128 combinations based on factor levels. Here, adeno, squamous, large, and small represent adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and small cell carcinoma, respectively.

<table>
<thead>
<tr>
<th>Group Name</th>
<th>Stage (X1)</th>
<th>Grade (X2)</th>
<th>Histology (X3)</th>
<th>Gender (X4)</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comb 1</td>
<td>I</td>
<td>1</td>
<td>adeno</td>
<td>1</td>
<td>1008</td>
</tr>
<tr>
<td>Comb 2</td>
<td>I</td>
<td>1</td>
<td>adeno</td>
<td>2</td>
<td>1426</td>
</tr>
<tr>
<td>Comb 3</td>
<td>I</td>
<td>1</td>
<td>squamous</td>
<td>1</td>
<td>430</td>
</tr>
<tr>
<td>Comb 4</td>
<td>I</td>
<td>1</td>
<td>squamous</td>
<td>2</td>
<td>187</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Comb 128</td>
<td>IV</td>
<td>4</td>
<td>small</td>
<td>2</td>
<td>3368</td>
</tr>
</tbody>
</table>

Table 3. Seven groups produced by cutting the dendrogram in Fig. 2 at the height 0.93.

### Results from Cluster Analysis

Before running our algorithm EACCD, we used the levels of four factors $X_1$, $X_2$, $X_3$, and $X_4$ to partition the dataset into $128(=4 \times 4 \times 4 \times 2)$ combinations, shown in Table 2. Due to the approximation of the chi-square distribution to the log-rank test statistic, a combination containing less than 100 patients was dropped from our study. In this case, no further analysis was done for these combinations, and our attention was paid to all the other combinations that have a size equal to or larger than 100. For example, Comb 5, as shown in Table 2, was dropped from our study. Under this restriction we only kept 80 combinations, leaving out a total of 1,264 patients. The output of cluster analysis for these 80 combinations is shown in Fig. 2, where for simplicity Comb has been removed from each combination or label.

### Prognostic System

It is straightforward to use the dendrogram shown in Fig. 2. Cutting off the dendrogram at a specified height of the dissimilarity axis partitions data into disjoint clusters or groups. Cutting at different heights usually leads to different numbers of groups. As an example, if we cut the dendrogram in Fig. 2 at a height slightly above 0.90, then we obtain 7 groups shown in Table 3. The log-rank test shows that any two groups differ significantly (using a significance level of 0.01) in their survival functions. Fig. 3 shows the Kaplan-Meier estimates ([6]) of the survival curves for the 7 groups. These 7 groups and their survival curves constitute a prognostic system for lung cancer patients, as discussed in step (ii) of the Section of Introduction. Prediction using this system is then carried out in the usual way. In comparison, those 4 survival curves from the TNM system, based on all the patients from the 80 combinations, are provided in Fig. 4.

### Conclusion

In this paper we have introduced a clustering based approach to establish prognostic systems that can be used to
Figure 2. Dendrogram from clustering of lung cancer data.
Survival Time in Months
Proportion Surviving

Kaplan-Meier Curves
Group 1
Group 2
Group 3
Group 4
Group 5
Group 6
Group 7

Figure 3. Survival curves of seven groups in Table 3.

Kaplan-Meier Curves
Stage I
Stage II
Stage III
Stage IV

Figure 4. Survival curves of four TNM stages.

predict an outcome or a survival rate of cancer patients. An application of the approach to lung cancer patients has been given.

Generalizing or refining the work presented in this paper can be done in many ways. Our algorithm EACCD actually is a two-stage clustering method. In the first stage, a dissimilarity measure is learnt by a partitioning method, and in the second stage, the learnt dissimilarity is used with a hierarchical clustering algorithm to obtain clusters of patients. These clusters of patients form a basis of a prognostic system. The effect of different algorithms used in each stage will be investigated in our future work. Refined algorithms, based on EACCD, will be sought and resulting prognostic systems with clinical applications will be reported. This constitutes our main research work in the future.

Acknowledgements

This work was partially supported by the National Science Foundation grant CCF-0729080.

References


