2011 11th IEEE International Conference on Data Mining Workshops

Identifying HotSpots in Lung Cancer Data Using Association Rule Mining

Ankit Agrawal and Alok Choudhary
Dept. of Electrical Engg. and Computer Science
Northwestern University
2145 Sheridan Rd, Evanston, IL 60201, USA
Email: {ankitag,choudhar}@eecs.northwestern.edu

***Abstract*—We analyze the lung cancer data available from the SEER program with the aim of identifying hotspots using association rule mining techniques. A subset of 13 patient attributes from the SEER data were recently linked with the survival outcome using prediction models, which is used in this study for segmentation. The goal here is to identify characteristics of patient segments where average survival is significantly higher/lower than average survival across the entire dataset. Automated association rule mining techniques resulted in hundreds of rules, from which many redundant rules were manually removed based on domain knowledge. The resulting rules conform with existing biomedical knowledge and provide interesting insights into lung cancer survival.**

***Keywords*-Association rule mining; hotspots; lung cancer;**

I. Introduction

Lung cancer ranks second in the list of most common cancers [1], and first in the list of most deadly cancers [2], with the survival rate being about 15% after 5 years of diagnosis [3].

The Surveillance, Epidemiology, and End Results (SEER) Program [4] of the National Cancer Institute is an authoritative repository of cancer statistics in the United States [5]. It is a population-based cancer registry which covers about 26% of the US population across several geographic regions and is the largest publicly available domestic cancer dataset. The data includes patient demographics, cancer type and site, stage, first course of treatment, and follow-up vital status. The SEER program collects cancer data for all invasive and in situ cancers, except basal and squamous cell carcinomas of the skin and in situ carcinomas of the uterine cervix [3]. The 'SEER limited-use data' is available from the SEER website on submitting a SEER limited-use data agreement form. [6] presents an overview study of the cancer data at all sites combined and on selected, frequently occurring cancers from the SEER data. The SEER data attributes can be broadly classified as demographic attributes (e.g. age, gender, location), diagnosis attributes (e.g. primary site, histology, grade, tumor size), treatment attributes (e.g. surgical procedure, radiation therapy), and outcome attributes (e.g. survival time, cause of death), which makes the SEER data ideal for performing outcome analysis studies.

Recently, lung cancer data from SEER was used to construct predictive models for lung cancer survival after 6 months, 9 months, 1 year, 2 years, and 5 years of diagnosis using several machine techniques [7]. In this work, we use same dataset with 13 predictor attributes as used for the lung cancer outcome calculator in [7] for association rule mining analysis.

The rest of the paper is organized as follows: Section II gives an overview of association rule mining and describes the HotSpot algorithm used in this work, followed by a description of the data and its attributes in Section III. Experiments and results are presented in Section IV, and finally the conclusion and future work in Section V.

II. Association Rule Mining

Association rule mining is often stated as follows [8]: Let I be a set of n binary attributes called items. Let T be a set of transactions. Each transaction in T contains a subset of the items in I. A rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \phi$. The sets of items X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively. A commonly given example from market basket analysis is of the rule $\{Bread\} \Rightarrow \{Butter\}$, meaning that customers who buy bread also buy butter.

Association rule mining is popularly done with flag attributes, indicating the presence/absence of the item in the transaction. However, even from nominal attributes (having multiple but finite possible values), and numeric attributes, it is possible to derive flag attributes for the purpose of association rule mining.

HotSpot Algorithm

This is an association rule mining algorithm which is directed by a target attribute, which means that the RHS or consequent is fixed to the target attribute. It can be used for segmentation with both nominal and numeric targets, where the LHS or antecedent would define the segment characteristics for segments which are significantly different from the entire dataset in terms of the target attribute. For example, if the target is a numeric attribute like the patient survival time, and the average patient survival time is t_{avg},

978-0-7695-4409-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDMW.2011.93

then it would be interesting to find segments in the data where the average patient survival time is higher/lower than t_{avg}. Similarly, if the target is a nominal attribute like 5-year-survival (whether or not a patient survived for at least 5 years), and the fraction of survived patients in the entire dataset is f, then it would be interesting to find segments where this fraction is higher/lower than f.

It uses a greedy approach to construct the tree of rules in a depth-first fashion, where the search is constrained by the following parameters:

1) **Maximum branching factor**: The number of children nodes to consider at each node. This parameter controls the amount of search performed, since the algorithm uses a greedy search.
2) **Minimum improvement in target value**: This is the minimum improvement in the target value of the resulting segment in order to consider adding a new branch.
3) **Minimum segment size**: The size of the resulting segment must be at least this much in order to add a new branch.

The HotSpot algorithm then, is straightforward. It begins with the entire dataset at the top, and goes down the data in a depth-first fashion using a greedy approach, i.e., it branches on that attribute which gives the maximum improvement in target value subject to the above constraints, and recursively tries the same at every node. Each node represents a segment, and hence, an association rule.

The improvement in the target value can be defined as either an increase or a decrease in the average target value (in case of numeric targets) or target fraction (in case of nominal targets).

We use the implementation of the HotSpot algorithm provided in the WEKA data mining toolkit [9].

III. Lung Cancer Data for Association Rule Mining

The lung cancer outcome calculator [7] mentioned earlier uses the data from the SEER November 2008 Limited-Use Data files [4] (released in April 2009) from nine SEER registries (Atlanta, Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco-Oakland, Seattle-Puget Sound, and Utah). This data had a follow-up cutoff date of December 31, 2006, i.e., the patients were diagnosed and followed-up upto this date. Subsequently, the data was selected for the patients diagnosed between 1998 and 2001, due to various reasons mentioned in [7], the primary being that it allowed predictive modeling of survival of upto 5-years (since the follow-up cutoff date was December 31, 2006, data used was for cancer patients with year of diagnosis as 2001 or before). The lung cancer outcome calculator uses the following 13 patient attributes:

1) **Age at diagnosis**: Numeric age of the patient at the time of diagnosis for lung cancer.
2) **Birth place**: The place of birth of the patient.
3) **Cancer grade**: A descriptor of how the cancer cells appear and how fast they may grow and spread.
4) **Diagnostic confirmation**: The best method used to confirm the presence of lung cancer.
5) **Farthest extension of tumor**: The farthest documented extension of tumor away from the lung, either by contiguous extension (regional growth) or distant metastases (cancer spreading to other organs far from primary site through bloodstream or lymphatic system).
6) **Lymph node involvement**: The highest specific lymph node chain that is involved by the tumor. Cancer cells can spread to lymph nodes near the lung, which are part of the lymphatic system (the system that produces, stores, and carries the infection-fighting-cells. This can often lead to metastases.
7) **Type of surgery performed**: The surgical procedure that removes and/or destroys cancerous tissue of the lung, performed as part of the initial work-up or first course of therapy.
8) **Reason for no surgery**: The reason why surgery was not performed (if not).
9) **Order of surgery and radiation therapy**: The order in which surgery and radiation therapies were administered for those patients who had both surgery and radiation.
10) **Scope of regional lymph node surgery**: It describes the removal, biopsy, or aspiration of regional lymph node(s) at the time of surgery of the primary site or during a separate surgical event.
11) **Cancer stage**: A descriptor of the extent the cancer has spread, taking into account the size of the tumor, depth of penetration, metastasis, etc.
12) **Number of malignant tumors in the past**: An integer denoting the number of malignant tumors in the patient's lifetime so far.
13) **Total regional lymph nodes examined**: An integer denoting the total number of regional lymph nodes that were removed and examined by the pathologist.

Table I presents the attributes of the lung cancer dataset, and Tables II-X present the possible values and codes of all the nominal attributes, except birth place, since there are too many possible values for birth place.

For association rule mining analysis, we removed all missing/unknown values, since we are interested in finding segments with precise definitions in terms of patient attributes. The survival time (in months) was chosen as the target attribute for the HotSpot algorithm. The dataset had 13,033 instances, 13 input patient attributes, and 1 target attribute. The average survival time in the entire dataset was 24.45 months. So, we would be interested to find segments of patients where the average survival time is significantly

Table I
LUNG CANCER DATASET ATTRIBUTES

Attribute	Type
Age at diagnosis	Numeric
Birth place	Nominal
Cancer grade	Nominal
Diagnostic confirmation	Nominal
Farthest extension of tumor	Nominal
Lymph node involvement	Nominal
Type of surgery performed	Nominal
Reason for no surgery	Nominal
Order of surgery and radiation therapy	Nominal
Scope of regional lymph node surgery	Nominal
Cancer stage	Nominal
Number of malignant tumors in the past	Numeric
Total regional lymph nodes examined	Numeric
Survival time	Numeric

Table II
CODES FOR CANCER GRADE

Code	Description
1	Grade I (well-differentiated)
2	Grade II (moderately differentiated)
3	Grade III (poorly differentiated)
4	Grade IV (undifferentiated)
9	Not determined/not stated/NA

Table III
CODES FOR DIAGNOSTIC CONFIRMATION

Code	Description
1	Positive histology
2	Positive cytology
4	Positive microscopic confirmation (method unspecified)
5	Positive laboratory test/marker study
6	Direct visualization without microscopic confirmation
7	Radiology and other imaging techniques without microscopic confirmation
8	Clinical diagnosis only (other than above)
9	Unknown whether microscopically confirmed

Table IV
CODES FOR TUMOR EXTENSION

Code	Description
0	In situ (Noninvasive/intraepithelial)
10	Tumor confined to one lung (excl. primary in MSB)
20	Main stem bronchus >2.0cm from carina
25	Primary confined to carina
30	Localized (NOS)
40	Pleura/Visceral/Pulmonary ligament without pleural effusion
50	Main stem bronchus <2.0cm from carina
60	Chest (thoriac) wall/Diaphragm
65	Separate tumor nodule(s) in the same lobe
70	Carina/Trachea/Esophagus
71	Heart/Visceral pericardium
72	Pleural effusion
73	Adjacent rib
75	Sternum/Vertebra(e)/Skeletal muscle/Skin of chest
77	Separate tumor nodule(s) in different lobe
78	Separate tumor nodule(s) in contralateral lung
79	Pericardial effusion
80	Further contiguous extension
85	Metastatis
99	Unknown if extension or metastatis

Table V
CODES FOR LYMPH NODE INVOLVEMENT

Code	Description
0	No lymph node involvement
1	Intrapulmonary/Hilar/Peribronchial
2	Subcarinal/Carinal/Mediastinal/ Tracheal/Aortic/Pulmonary ligament/Pericardial
5	Regional lymph node(s)
6	Contralateral hilar/Supraclavicular/ Ipsilateral/Contralateral/ Scale
7	Other than above (incl. cervical neck nodes)
8	Distant lymph nodes (NOS)
9	Unknown/Not stated

Table VI
CODES FOR TYPE OF SURGERY

Code	Description
0	No surgery
12	Laser ablation/cryosurgery
13	Electrocautery/Fulguration
15	Local tumor destruction (NOS)
19	Local tumor destruction or excision (NOS)
20	Excision or resection of less than one lobe (NOS)
21	Wedge resection
22	Segmental resection (including lingulectomy)
23	Excision (NOS)
24	Laser excision
25	Bronchial sleeve resection only
30	Resection of lobe or bilobectomy, but less than whole lung
33	Lobectomy with mediastinal lymph node dissection
45	Lobe or bilobectomy extended (NOS)
46	Lobe or bilobectomy extended with chest wall
55	Pneumonectomy (NOS)
56	Pneumonectomy with mediastinal lymph node dissection
65	Extended pneumonectomy
66	Extended pneumonectomy plus pleura or diaphragm
70	Extended radical pneumonectomy
80	Resection of lung (NOS)
90	Surgery (NOS)
99	Unknown if surgery performed

Table VII
CODES FOR REASON FOR NO SURGERY

Code	Description
0	Surgery performed
1	Surgery not recommended
2	Contraindicated due to other conditions
6	Unknown reason for no surgery
7	Patient or patient's guardian refused
8	Recommended, unknown if done
9	Unknown if surgery performed

Table VIII
CODES FOR ORDER OF SURGERY AND RADIATION THERAPY

Code	Description
0	No radiation and/or surgery
2	Radiation before surgery
3	Radiation after surgery
4	Radiation both before and after surgery
5	Intraoperative radiation therapy
6	Intraoperative radiation with other radiation given before/after surgery
9	Sequence unknown, but both surgery and radiation were given

Table IX
CODES FOR SCOPE OF REGIONAL LYMPH NODE SURGERY

Code	Description
0	No regional lymph nodes removed
1	Regional lymph nodes removed (NOS)
2	Intrapulmonary/ipsilateral hilar/ipsilateral peribronchial nodes
3	Ipsilateral mediastinal/subcarinal nodes
4	Combination of 2 and 3
5	Contralateral mediastinal/contralateral hilar/ipsilateral/contralateral scalene/supraclavicular nodes
6	Combination of 5 with 2 or 3
9	Unknown/not stated

Table X
CODES FOR CANCER STAGE

Code	Description
0	In situ (Noninvasive neoplasm)
1	Localized (Invasive neoplasm confined to the lung)
2	Regional (Extended neoplasm)
7	Distant (Spread neoplasm)
9	Unstaged/Unknown

higher than and lower than 24.45 months.

IV. EXPERIMENTS AND RESULTS

The distribution of survival time across all the patients is shown in Figure 1. Before performing HotSpot analysis, we would like to study the influence of each of the individual 13 attributes on survival time. For this purpose, we plotted the average survival time for different possible values of each input attribute, as shown in Figures 2-14.

We performed two independent analyses to find segments in which average survival time was higher and lower than overall average survival. Several combinations of algorithm parameters (maximum branching factor, minimum improvement in target value, and minimum segment size) were tried. Here we report the results with the following parameters: maximum branching factor = 3, minimum improvement in target value = 1%, and minimum segment size = 100.

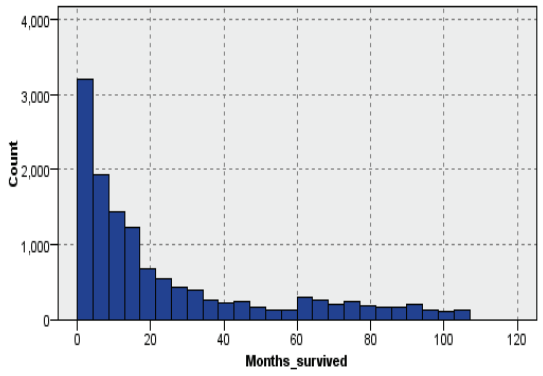

Figure 1. Distribution of survival time (in months).

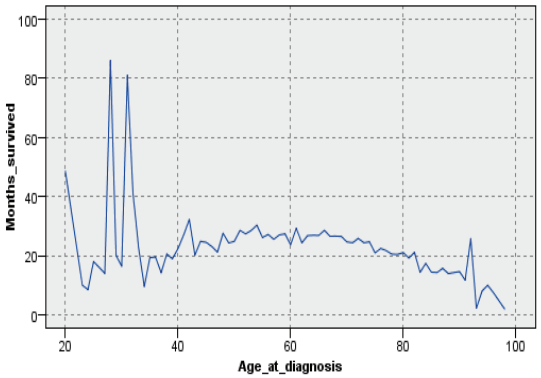

Figure 2. Survival time vs. Age at diagnosis.

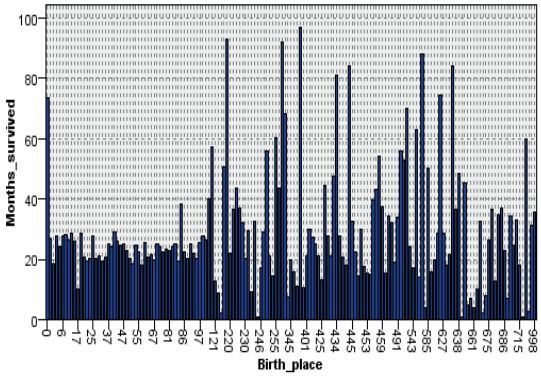

Figure 3. Survival time vs. Birth place.

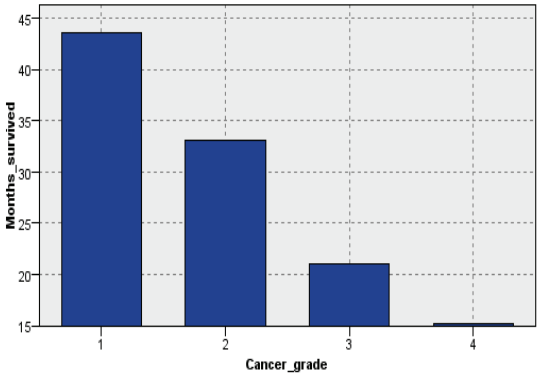

Figure 4. Survival time vs. Cancer grade.

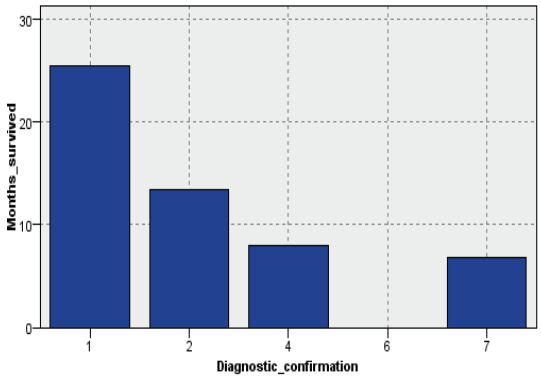

Figure 5. Survival time vs. Diagnostic confirmation.

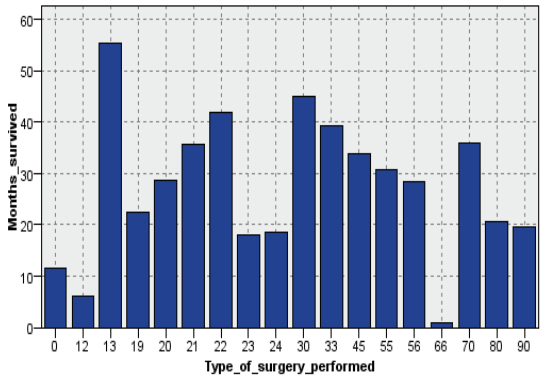

Figure 8. Survival time vs. Type of surgery performed.

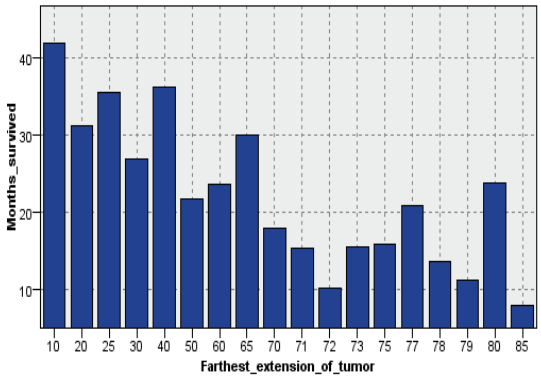

Figure 6. Survival time vs. Farthest extension of tumor.

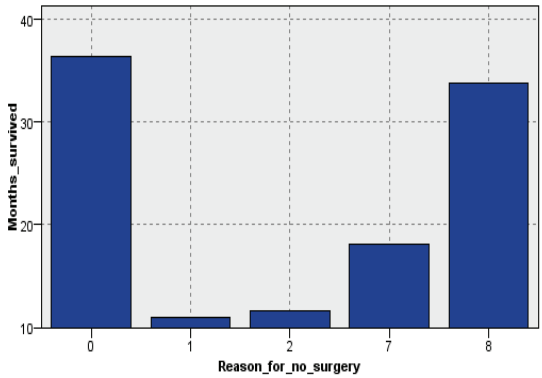

Figure 9. Survival time vs. Reason for no surgery.

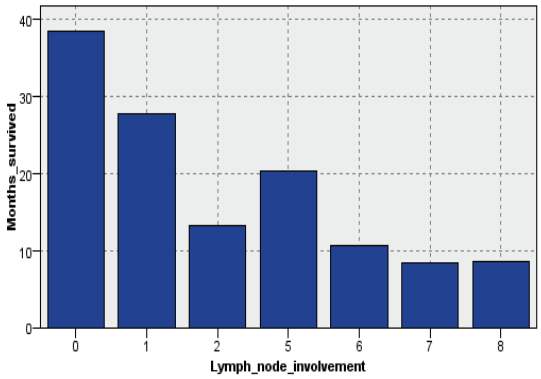

Figure 7. Survival time vs. Lymph node involvement.

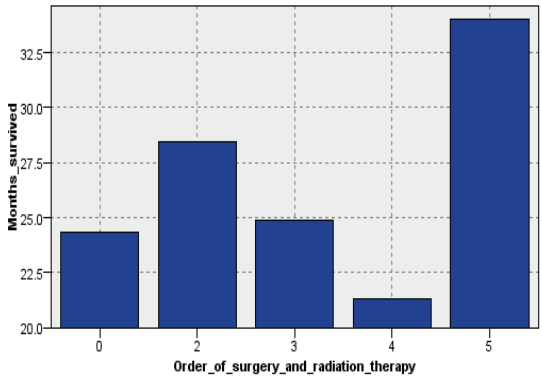

Figure 10. Survival time vs. Order of surgery and radiation therapy.

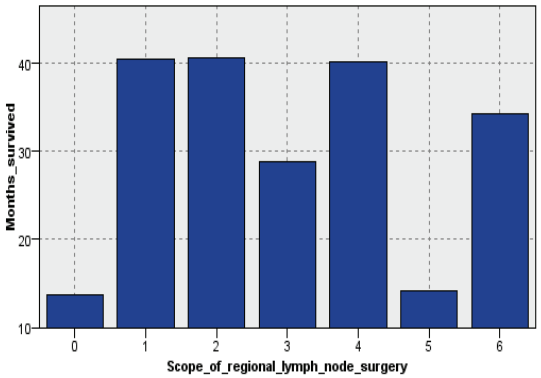

Figure 11. Survival time vs. Scope of regional lymph node surgery.

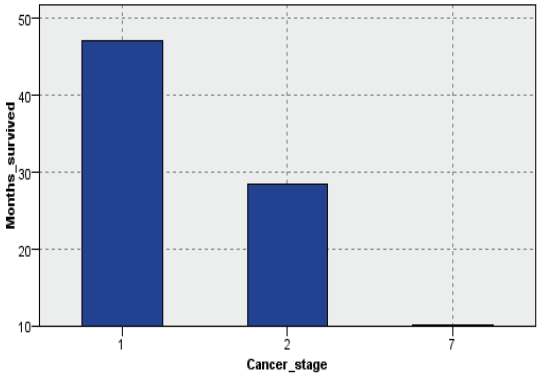

Figure 12. Survival time vs. Cancer stage.

Figure 13. Survival time vs. Number of malignant tumors in the past.

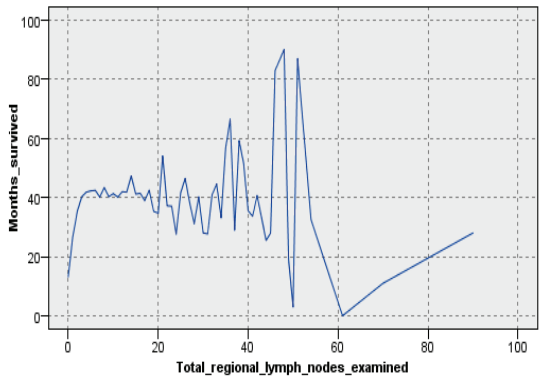

Figure 14. Survival time vs. Total regional lymph nodes examined.

Table XI
NUMBER OF ASSOCIATION RULES

Mode	Generated by HotSpot Algorithm	Redundant Rule Removal - Stage I	Redundant Rule Removal - Stage II
Higher target value	165	85	12
Lower target value	64	32	6

As commonly known, association rule analysis invariably leads to discovery of a large number of redundant rules, which need to be removed. We used a 2-stage semi-manual procedure to remove redundant rules:

1) **Stage I**: Since the HotSpot algorithm tries to go deeper into the data as long as it is able to improve the target value, the leaf nodes would have the best target value compared to all the nodes on its path. So, we discard all the rules corresponding to the non-leaf nodes, and retain only the rules corresponding to the leaf nodes. This stage does not require manual intervention.
2) **Stage II**: Even after Stage I, there still remain quite a few redundant rules, the removal of which require domain expertise. The redundant rules at this stage were manually removed.

Table XI present the number of rules generated by the HotSpot algorithm, and the rules after each stage of redundant rule removal.

Lift of a rule is the relative improvement in the target (here survival time) as compared to the average value of the target across the entire dataset. Therefore, lift for the two modes can be defined as follows:

$$Lift_{higher} = Avg.SurvivalTimeWithinSegment/24.45$$
$$Lift_{lower} = 24.45/Avg.SurvivalTimeWithinSegment$$

Table XII
NON-REDUNDANT ASSOCIATION RULES DENOTING SEGMENTS WHERE AVERAGE SURVIVAL TIME IS SIGNIFICANTLY HIGHER THAN 24.45 MONTHS

Non-redundant rules	Avg. Survival time	Segment size	Lift	Segment description
Cancer stage = 1, Cancer grade = 1, Total regional lymph nodes examined > 3, Order of surgery and radiation therapy = 0, Type of surgery performed = 30, Age at diagnosis <= 78, Number of malignant tumors in lifetime <= 2, Total regional lymph nodes examined <= 17	68.18	100	2.788559462	The tumor is well-differentiated and localized, regional lymph nodes examined is between 4 and 17, age of the patient at time of diagnosis is less than 79, current tumor is patient's first or second tumor, and resection of lobe/bilobectomy is performed by the surgeon.
Cancer stage = 1, Age at diagnosis <= 52, Type of surgery performed = 30, Total regional lymph nodes examined <= 14, Total regional lymph nodes examined > 0, Age at diagnosis > 38	68.11	100	2.785696465	The tumor is localized, age of patient is between 39 and 52, number of regional lymph nodes examined is between 1 and 14, and resection of lobe/bilobectomy is performed by the surgeon.
Type of surgery performed = 30, Cancer grade = 1, Scope of regional lymph node surgery = 1, Total regional lymph nodes examined <= 14	66.8317	101	2.733414043	Tumor is well-differentiated, number of regional lymph nodes examined is less than 15, resection of lobe/bilobectomy is performed, and regional lymph nodes are removed.
Cancer stage = 1, Age at diagnosis <= 52, Type of surgery performed = 30, Farthest extension of tumor = 10, Age at diagnosis > 40	66.2613	111	2.710084704	Tumor is localized, age of patient is between 41 and 52, tumor is confined to one lung, and resection of lobe/bilobectomy is performed.
Type of surgery performed = 30, Birth place = 99, Lymph node involvement = 0, Age at diagnosis <= 75	64.9811	106	2.657724571	Patient was born in Hawaii, patient's age is less than 76, there is no lymph node involvement, and resection of lobe/bilobectomy is performed.
Cancer stage = 1, Birth place = 99, Reason for no surgery = 0, Age at diagnosis <= 82	63.9604	101	2.615977979	Tumor is localized, patient was born in Hawaii, patient's age is less than 83, and surgery was performed.
Cancer grade = 1, Total regional lymph nodes examined > 6, Lymph node involvement = 0, Age at diagnosis <= 80, Total regional lymph nodes examined <= 18	63.8614	101	2.611928883	Tumor is well-diffentiated, number of lymph nodes examined is between 7 and 18, there is no lymph node involvement, and patient's age is less than 81.
Type of surgery performed = 30, Cancer stage = 1, Birth place = 7, Farthest extension of tumor = 10, Total regional lymph nodes examined > 2	63.0971	103	2.580669042	Tumor is localized, patient was born in Connecticut, tumor is confined to one lung, number of lymph nodes examined is greater than 2, and resection of lobe/bilobectomy is performed.
Cancer grade = 1, Scope of regional lymph node surgery = 2, Lymph node involvement = 0, Age at diagnosis <= 75	62.16	100	2.542341686	Tumor is well-differentiated, there is no lymph node involvement, patient's age is less than 76, and intrapulmonary/ipsilateral hilar/ipsilateral peribronchial nodes are removed.
Cancer stage = 1, Birth place = 99, Farthest extension of tumor = 10, Age at diagnosis <= 81	60.3762	101	2.469384333	Tumor is localized (confined to one lung), patient is born in Hawaii and is less than 82 years old.
Cancer stage = 1, Birth place = 99, Farthest extension of tumor = 10, Diagnostic confirmation = 1	60.1845	103	2.46154381	Tumor is localized (confined to one lung), patient is born in Hawaii, and cancer was confirmed by positive histology.
Type of surgery performed = 30, Cancer stage = 1, Birth place = 97	58.71	100	2.401236815	Tumor is localized, patient is born in California, and resection of lobe/bilobectomy is performed by the surgeon.

Tables XII and XIII present the non-redundant association rules obtained with 'higher' and 'lower' mode respectively. The description of the segment features is also included in the tables.

Most of the rules obtained in both cases conform with existing biomedical knowledge and provide interesting insights into lung cancer survival.

V. CONCLUSION AND FUTURE WORK

In this paper, we performed association rule mining analysis on lung cancer data from SEER to identify hotspots in the cancer data, where the patient survival time is significantly higher than and lower than the average survival time across the entire dataset.

We believe that such analysis can be very useful to identify the factors affecting survival, and aid doctors and patients in avoiding the conditions which are known to reduce survival time, and encourage the conditions which are known to increase the survival time, whenever possible. It can also aid doctors in decision making and improve informed patient consent by providing a better understanding of the risks involved in a particular treatment procedure.

Similar analysis can also be done for other cancers.

ACKNOWLEDGMENT

We thank the SEER program to make their limited-use data available for this work. This work is supported in part by NSF award numbers CCF-0621443, OCI-0724599, CCF-0833131, CNS-0830927, IIS-0905205, OCI-0956311, CCF-0938000, CCF-1043085, CCF-1029166, and OCI-1144061, and in part by DOE grants DE-FC02-07ER25808, DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, and DE-SC0005340.

REFERENCES

[1] "Introduction to lung cancer," national Cancer Institute, SEER training modules, URL: http://training.seer.cancer.gov/lung/intro/ accessed: Aug 2, 2011.

Table XIII
NON-REDUNDANT ASSOCIATION RULES DENOTING SEGMENTS WHERE AVERAGE SURVIVAL TIME IS SIGNIFICANTLY LOWER THAN 24.45 MONTHS

Non-redundant rules	Avg. Survival time	Segment size	Lift	Segment description
Farthest extension of tumor = 85, Lymph node involvement = 7, Order of surgery and radiation therapy = 0, Scope of regional lymph node surgery = 0, Cancer grade = 3	5.21	100	4.692879079	Tumor has metastasized and is poorly differentiated, lymph nodes are involved in metastasis, and no lymph nodes are removed.
Farthest extension of tumor = 85, Birth place = 99, Order of surgery and radiation therapy = 0, Type of surgery performed = 0, Cancer grade = 3	5.6727	110	4.310099247	Tumor has metastasized and is poorly differentiated, no surgery was performed, and the patient was born in Hawaii.
Farthest extension of tumor = 85, Birth place = 99, Order of surgery and radiation therapy = 0, Type of surgery performed = 0, Diagnostic confirmation = 1	5.7344	128	4.263724191	Tumor has metastasized, no surgery was performed, cancer was confirmed by positive histology, and patient was born in Hawaii.
Farthest extension of tumor = 85, Reason for no surgery = 2, Diagnostic confirmation = 1	5.7803	132	4.229866962	Tumor has metastasized, surgery was contraindicated and not performed, and cancer was confirmed by positive histology.
Farthest extension of tumor = 72, Cancer grade = 3, Type of surgery performed = 0, Lymph node involvement = 2, Order of surgery and radiation therapy = 0	7.5268	205	3.248379125	Pleural effusion has taken place, tumor is poorly differentiated, subcarinal/carinal/mediastinal/tracheal/aortic/ pulmonary ligament/pericardial lymph nodes are involved, and no surgery was performed.
Farthest extension of tumor = 72, Diagnostic confirmation = 2, Reason for no surgery = 1	8.5982	112	2.843606801	Pleural effusion has taken place, cancer was confirmed by positive cytology, surgery was not recommended and hence not performed.

[2] "Lung cancer statistics," centers for Disease Control and Prevention, URL: http://www.cdc.gov/cancer/lung/statistics/ accessed: Aug 2, 2011.

[3] L. A. G. Ries and M. P. Eisner, *Cancer of the lung*. National Cancer Institute, SEER Program, 2007, ch. 9.

[4] "Surveillance, epidemiology, and end results (seer) program (www.seer.cancer.gov) limited-use data (1973-2006)," National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, 2008, released April 2009, based on the November 2008 submission.

[5] "Overview of the seer program," surveillance Epidemiology and End Results, URL: http://seer.cancer.gov/about/ accessed: Aug 2, 2011.

[6] L. A. Gloeckler Ries, M. E. Reichman, D. R. Lewis, B. F. Hankey, and B. K. Edwards, "Cancer Survival and Incidence from the Surveillance, Epidemiology, and End Results (SEER) Program," *Oncologist*, vol. 8, no. 6, pp. 541–552.

[7] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "A lung cancer outcome calculator using ensemble data mining on seer data," in *Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics*, ser. BIOKDD '11, 2011, pp. 5:1–5:9.

[8] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, ser. SIGMOD '93, 1993.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.