A Framework for Improving Breast Cancer Care Decisions by using Self-Organizing Maps to Profile Patients and Quantify their Attributes

Vanda Victoria Spencer

Follow this and additional works at: https://scholarsjunction.msstate.edu/td

Recommended Citation
Spencer, Vanda Victoria, "A Framework for Improving Breast Cancer Care Decisions by using Self-Organizing Maps to Profile Patients and Quantify their Attributes" (2018). Theses and Dissertations. 133. https://scholarsjunction.msstate.edu/td/133

This Graduate Thesis is brought to you for free and open access by the Theses and Dissertations at Scholars Junction. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholars Junction. For more information, please contact scholcomm@msstate.libanswers.com.
A framework for improving breast cancer care decisions by using self-organizing maps to profile patients and quantify their attributes

By

Vanda Victoria Spencer

A Thesis
Submitted to the Faculty of Mississippi State University
in Partial Fulfillment of the Requirements for the Degree of Master of Science in Industrial and Systems Engineering in the Bagley College of Engineering

Mississippi State, Mississippi
August 2018
A framework for improving breast cancer care decisions by using self-organizing maps to profile patients and quantify their attributes

By

Vanda Victoria Spencer

Approved:

______________________
Brian Smith
(Major Professor)

______________________
Mohammad Marufuzzaman
(Committee Member)

______________________
Michael Andre Hamilton
(Committee Member)

______________________
Stanley F. Bullington
(Graduate Coordinator)

______________________
Jason M. Keith
Dean
Bagley College of Engineering
Date of Degree: August 10, 2018

Institution: Mississippi State University

Major Field: Industrial and Systems Engineering

Select Appropriate Title: Brian K. Smith

Title of Study: A framework for improving breast cancer care decisions by using self-organizing maps to profile patients and quantify their attributes

Pages in Study 51

Candidate for Degree of Master of Science

Considering the commonality of breast cancer among women in the United States and the increasing popularity of precision medicine and data analytics in healthcare, the aim of this study was to use self-organizing maps (SOM) to profile and make decisions about breast cancer patients. Breast cancer mass measurements were combined with nine non-medical attributes—family income, history of cancer, level of education, preference of probability level, presence of dependents, employment status, marital status, age, and location—that were randomly generated based on recent population statistics and fed into a SOM. The SOM’s accuracy was evaluated at around 80%. To show the decision-making capabilities of the SOM, a subset of the patients were treated as new patients and placed on the map. Profiles of these clusters were created to show how decisions made about patients’ diagnosis, delivery, and treatment differed based on the cluster to which they belonged.
DEDICATION

I would like to dedicate this thesis to my parents, whose constant love and support are the basis for all my past, current, and future successes.
ACKNOWLEDGEMENTS

I would like to express my deepest thanks to Roy Jafari-Marandi, Ph.D. His willingness to provide time and guidance and his encouraging attitude instilled me with the confidence to complete this work.

Additionally, I would like to thank my advisor and committee chair, Brian Smith, Ph.D. His moral support and involvement from the beginning made this thesis possible.
# TABLE OF CONTENTS

DEDICATION.......................................................................................................................... ii  

ACKNOWLEDGEMENTS........................................................................................................... iii  

LIST OF TABLES................................................................................................................... vi  

LIST OF FIGURES .................................................................................................................. vii  

CHAPTER  

I. INTRODUCTION .................................................................................................................. 1  

II. LITERATURE REVIEW ....................................................................................................... 2  

   2.1 Self-Organizing Maps ................................................................................................. 2  
   2.2 Error Types and Costs ................................................................................................. 3  
   2.3 Unconventional Decision-Making .............................................................................. 4  
   2.4 Decision-Making in Case of Breast Cancer ............................................................... 6  
   2.5 Precision Medicine .................................................................................................... 7  

III. METHODOLOGY ............................................................................................................... 8  

   3.1 Dataset .......................................................................................................................... 8  
   3.2 Data Preparation .......................................................................................................... 9  
      3.2.2 Family Income ..................................................................................................... 11  
      3.2.3 Cancer History ................................................................................................... 11  
      3.2.4 Preference ............................................................................................................ 12  
      3.2.5 Presence of Dependents ....................................................................................... 12  
      3.2.6 Level of Education and Employment Status ....................................................... 12  
      3.2.7 Marital Status ...................................................................................................... 13  
      3.2.8 Age ....................................................................................................................... 14  
      3.2.9 Location ................................................................................................................ 14  
   3.3 Randomized Data Generation ...................................................................................... 15  
   3.4 SOM Construction ....................................................................................................... 15  

IV. ANALYSIS .......................................................................................................................... 17  

V. APPLICATIONS .................................................................................................................. 21  

iv
VI. LIMITATIONS AND FURTHER RESEARCH..........................................................28
REFERENCES ........................................................................................................29

APPENDIX
A. NON-MEDICAL DATA GENERATION CODE .................................................32
B. SOM CODE ......................................................................................................40
C. NON-MEDICAL ATTRIBUTE MEAN MAPS ....................................................46
# LIST OF TABLES

3.1 Non-medical attribute generation values...................................................... 10
5.2 Cluster class evaluation .............................................................................. 23
5.3 Test patient cluster means for non-categorical attributes........................... 24
5.4 Test patient cluster percent mixes for categorical attributes ....................... 25
LIST OF FIGURES

4.1 SOM train hits map .................................................................................17
4.2 Weight plots for WDBC medical attributes ...........................................18
4.3 Weight plots for generated non-medical attributes .................................19
4.4 Mean label map .....................................................................................20
5.1 Test set hit map.......................................................................................21
5.2 Superimposed test set hit map and mean label map ..................................22
A.1 Age, preference, and location generation ...............................................33
A.2 Marital status and education level generation .........................................34
A.3 Employment status generation ...............................................................35
A.4 Presence of dependents generation .......................................................36
A.5 Presence of dependents (cont.) and cancer history generation ................37
A.6 Cancer history generation (cont.) .........................................................38
A.7 Family income generation ..................................................................39
B.1 Randomizing order, assigning attributes, transforming values, and creating train set .................................................................41
B.2 Initializing, training, and plotting SOM, and preparing heat map ............42
B.3 Preparing for heat map (cont.) ...............................................................43
B.4 Preparing heat map (cont.) ..................................................................44
B.5 Creating heat map and plotting test data .................................................45
C.1 Mean map for age .................................................................................47
C.2 Mean map for location ..........................................................................48
C.3 Mean map for preference level .............................................................. 48
C.4 Mean map for marital status ................................................................. 49
C.5 Mean map for education level ............................................................... 49
C.6 Mean map for employment status ......................................................... 50
C.7 Mean map for presence of dependents ................................................ 50
C.8 Mean map for history of cancer ............................................................ 51
C.9 Mean map for family income ................................................................. 51
CHAPTER I
INTRODUCTION

Artificial Neural Networks (ANNs) and Self-Organizing Maps (SOMs), among other data analysis tools, have aided in advancing the field of medicine by employing algorithms for data clustering and classification. These methods have been used to identify, characterize, and classify potentially cancerous breast masses as either benign or malignant (Chen, Chang, & Huang, 2000; Markey, Lo, Tourassi, & Floyd, 2003). While these methods have proven quite effective, they have incorporated only information from measurements and images concerning the mass (e.g. sonographic images and features; Chen et al., 2000; B. Zheng, Yoon, & Lam, 2014). Previous studies have neglected the decision-making capabilities of SOMs, which can be useful for including the thoughts of the doctors making the diagnoses and treatment plans and the feelings and lifestyles of the patients receiving them. This study attempts to address those shortcomings by focusing on the decision-making capabilities rather than diagnostic accuracy and incorporating patient and doctor data to tailor each diagnosis and treatment plan to those involved.
2.1 Self-Organizing Maps

A self-organizing map (SOM) is a type of “artificial neural network, the cells of which become specifically tuned to various input signal patterns or classes of patterns through an unsupervised learning process” (Kohonen, 1990) and was developed by Tuevo Kohonen; the following is Kohonen’s explanation of the maps and how they are produced. Used for clustering, classification, and predictions of datasets, the SOM is a valuable data visualization tool that uses an algorithm to display high-dimensional data on a low-dimensional map. The map consists of both input and output nodes with each input node connected to every output node. The Euclidean distance between a data point and each output node is calculated. Each data point is placed near the output node with the smallest Euclidean distance and then maintains its topological position on the map. Therefore, observations with similar outcomes and attributes are grouped together into cells, resulting in a map displaying similar cells near one another (Jafari-Marandi, Khanzadeh, Smith, & Bian, 2017).

Self-organizing maps have been used in myriad different fields and cases from predicting churn rate (Jafari-Marandi, Denton, Smith, & Keramati, 2017) to detecting cancer (Chen et al., 2000). A large portion of the medical uses focus on breast cancer in particular because breast cancer is so prevalent, affecting around 12% of women in the
United States (“U.S. Breast Cancer Statistics,” 2018). Breast cancer data has been the basis for numerous studies, including conducting a classification analysis (Markey et al., 2003), diagnosing breast cancer (Chen et al., 2000), and even using these diagnoses to attempt to reduce the need for breast biopsies (Y. Zheng, Greenleaf, & Gisvold, 1997). As the use of self-organizing maps is relatively new, especially to the field of healthcare, the main focus of many studies has been to improve the accuracy of SOMs predictive capabilities (Chen et al., 2000; Jafari-Marandi, Khanzadeh, et al., 2017). Chen et al.’s study achieved a breast cancer diagnostic accuracy of 85.6% (Chen et al., 2000). Though adjusting for accuracy is important, the other capabilities SOMs possess have yet to be evaluated. The use of SOMs as decision aids rather than just classification and diagnostic methods needs to be further explored.

2.2 Error Types and Costs

As with any method of statistical analysis, there exists some risk of error, manifesting in the field of healthcare research as either Type I (false positive) or Type II (false negative). In this field, a false positive refers to a test result stating some condition is present when it is not; a false negative refers to a test result stating some condition is not present when it is. Most general classifications operate under the assumption that these two types of errors are received the same way. However, as these classification methods are applied in fields where the human element is the primary concern, this assumption begins to shift. An area often left untouched is the impact of these two different errors, especially in the medical field where the diagnosis can mean life or death, as with cancer (Beins, 2017). When diagnosing breast cancer, a Type I error indicates a detection of cancer that was not actually present; a Type II error indicates
cancer was present but went undetected. As these errors are different, they carry different implications. The severity of the error often depends on a number of patient factors, such as age, stage of life, and personality type (Beins, 2017). In some cases, a false negative is viewed as more detrimental because the cancer goes untreated and patient health is put at risk. However, the costs associated with false positives can be just as life-altering. In addition to psychological stress, there exists a level of financial distress. With a cancer diagnosis comes additional costly testing. Beins 2017 concluded that around $4 billion dollars per year is spent on additional testing after a false positive diagnosis (Beins, 2017). Understanding these errors and their associated costs allows for a more tailored adjustment of the SOM.

2.3 Unconventional Decision-Making

While many studies in this area tend to focus on the use of algorithms like self-organizing maps to aid in more accurate diagnoses (Chen et al., 2000; Markey et al., 2003), some studies place an emphasis on more unconventional methods and decision-making (Restivo et al., 2016; Schilli, Stricklin, Payne, Rader, & Stoecker, 2014; Stricklin, Payne, Rader, Schilli, & Stoecker, 2014). The goal of one such study conducted by Restivo et al. was to determine if non-medical patient information affected the decisions made about treatment by multidisciplinary oncological teams. The decisions made (on a case-by-case basis) included whether to continue treatment and what type of treatment with which to proceed, though the decision was deferred in some cases to a later meeting. The non-medical patient factors mentioned in the meetings were coded and categorized by sociodemographic, psychological, and relational characteristics. A positive correlation
was found between the mention of non-medical factors and deferring the final decision, indicating these factors do affect decision-making (Restivo et al., 2016).

Another way unconventional methods have been used to aid in decision making is through hand examinations. One study (Schilli et al., 2014; Stricklin et al., 2014) used these physical examinations to reveal information about the daily activities, habits, and personality traits of patients and yielded the following findings: the location and type of discoloration, scars, and other injuries indicated certain psychological and physical states. For example, circumferential scars on the wrist can indicate suicide attempts, while scars on the knuckles can be a sign of bulimia. A patient may be reluctant to share this information, feeling it is embarrassing or irrelevant; however, this information could affect the decisions made about a diagnosis or treatment. Another aspect of this study indicated that the ratio of the length of the second digit (2D) to the fourth digit (4D) revealed personality aspects in men. A low 2D:4D ratio in males was found to indicate increased spatial and physical ability, while a high 2D:4D ratio indicated a detail-oriented nature and skills often found in computer programmers. Having this kind of information can allow doctors to tailor both the delivery of the treatment plan and the plan itself based on patient-specific attributes (Schilli et al., 2014).

This kind of decision-making, while unconventional, has its place. These cases not only show that different approaches to decision-making are valid and prove useful, but also that there exists room for new ways of interpreting data and algorithms that have long been used for only one purpose.
2.4 Decision-Making in Case of Breast Cancer

Decisions must be made at all stages of the breast cancer process, from prevention to diagnosis and treatment. Often the decision-making aspect of treatment is performed primarily by the doctor. However, some studies have begun to discuss the patients’ involvement. A study conducted Yong Hui Nies et al. in Malaysia attempted to find a connection between patients’ sociodemographic status (ethnicity, education, etc.) and their preferred role in the decision-making process. The study found that those with higher levels of education—diplomas, degrees, and post-graduate qualifications—were much more (7.52 times) likely to prefer an major role in making the decisions about their treatment (Yong Hui Nies et al., 2017). An additional case found that, in retrospect, most women studied regretted not playing a bigger decisional role in their treatment plan (Hack, Degner, Watson, & Sinha, 2006). The importance of including patients in the decision-making process to enhance patient satisfaction has become increasingly evident. While a few studies have been conducted to focus on the diagnostics and patient reception of their disease (National Cancer Center, Korea, 2011), most have concentrated on the later stages. In these cases, the decision-making aspect of choosing treatments and surgical options is the primary source of concern. The National Institutes of Health & Agency for Healthcare Research and Quality conducted a study where the surgical decision-making process for young women with breast cancer was evaluated to gain a better understanding of how to reduce treatment uncertainty, improve communications, and decrease patient anxiety (National Institutes of Health (NIH) & Agency for Healthcare Research and Quality (AHRQ), 2017). Tamirisa et al. (2017) interviewed both doctors and patients, finding that part of the communication issue between them stems
from a gap in knowledge of patient history (Tamirisa et al., 2017). One patient stated “No one ever asked me what my background was so they could know what level of information I could receive” (Tamirisa et al., 2017), showing different education levels and backgrounds are grounds for different information delivery tactics.

2.5 Precision Medicine

Precision medicine is a relatively new way of approaching medicine that allows it to be more personalized (i.e. blood typing). This movement is headed in the short-term direction of improving cancer prevention, diagnosis, and treatment with a more long-term goal of addressing a broader range of health issues and diseases (Collins & Varmus, 2015). People come from a variety of backgrounds and possess different genetics and experience different environments. Researchers acknowledge this can cause different reactions to certain diseases and medicines. Kennedy (2018) explained the work of looking at an individual’s genes and environment to aid in prevention and early detection (Kennedy, 2018). As precision medicine has been gaining popularity, studies have primarily focused on the treatment aspect, like targeted drug therapies (Samimi et al., 2017). However, it is important to consider the role of personalized diagnostics.
CHAPTER III
METHODOLOGY

3.1 Dataset

The Wisconsin Diagnostic Breast Cancer (WDBC) database from the University of California, Irvine, Machine Learning Repository was used as the basis for this study (Dua & E., 2017). The features in this database were computed from a digital image of a fine needle aspiration biopsy (FNAB) of a breast mass. A FNAB describes when a needle is used to draw fluid from a lump in the breast and the cells’ nuclei are then examined for cancer. The features describe characteristics of the nuclei in the FNAB image. The multivariate dataset contains 569 cases, each with 32 attributes. The 32 attributes consist of a case identification number, a diagnosis of the mass as benign (B) or malignant (M), and ten distinct measures, each having a mean, standard error, and the mean of the three largest measures for each attribute (resulting in 3 separate values for each measure). The ten distinct input features are: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. To reduce any data dependencies among the statistical measures, the standard error and mean of the three largest measures for each attribute were removed, leaving 12 distinct attributes for each case.
3.2 Data Preparation

To demonstrate the personalization and decision-making power of SOMs, the data from the WDBC database was supplemented with non-medical patient attributes expected to affect both information pertaining to and delivery of diagnoses and treatments. The nine additional attributes added to the dataset are as follows: family income, history of cancer, level of education, preference of probability level (the level at which the patient wants to be informed of the probability of the mass being malignant), presence of dependents, employment status, marital status, age, and location. To aid in creating the framework for this study, the data for each of these additional attributes was randomly generated in Matlab based on current United States population statistics from the United States Census Bureau.

To provide a larger dataset for the SOM tool to analyze, each of the 569 patients’ and their original measured attributes were repeated 100 times, resulting in 56,900 attributes. The nine supplemental attributes were randomized across these 56,900 cases.

As detailed in Table 3.1, arbitrary categorical values were assigned to the following attributes: history of cancer, level of education, preference of probability level, presence of dependents, employment status, marital status, and location. Family income and age were randomized within ranges. Because most patients diagnosed with breast cancer are women over the age of twenty, these random values were based on the U.S. female population over the age of twenty.
<table>
<thead>
<tr>
<th>Attribute</th>
<th>Quantification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family Income</td>
<td>Value from 0 to $400,000</td>
</tr>
<tr>
<td>Cancer History</td>
<td>0 – No family history of lung, colorectal, breast, ovarian, or prostate cancers</td>
</tr>
<tr>
<td></td>
<td>1 – Family history of lung, colorectal, breast, ovarian, or prostate cancers</td>
</tr>
<tr>
<td>Education Level</td>
<td>0 – No high school diploma / General Education Development (GED)</td>
</tr>
<tr>
<td></td>
<td>1 – High school diploma / GED</td>
</tr>
<tr>
<td></td>
<td>2 – Some college / Associates degree</td>
</tr>
<tr>
<td></td>
<td>3 – College degree (Bachelors, Masters, PhD, Professional degree)</td>
</tr>
<tr>
<td>Preference</td>
<td>0 – Desire to know if probability of cancer is over 50%</td>
</tr>
<tr>
<td></td>
<td>1 – Desire to know if probability of cancer is 30-50%</td>
</tr>
<tr>
<td></td>
<td>2 – Desire to know if probability of cancer is 10-30%</td>
</tr>
<tr>
<td></td>
<td>3 – Desire to know if probability of cancer is below 10%</td>
</tr>
<tr>
<td>Presence of Dependents</td>
<td>0 – No dependents present</td>
</tr>
<tr>
<td></td>
<td>1 – Dependents present</td>
</tr>
<tr>
<td>Employment Status</td>
<td>0 – Not participating in labor force</td>
</tr>
<tr>
<td></td>
<td>1 – Unemployed</td>
</tr>
<tr>
<td></td>
<td>2 – Employed</td>
</tr>
<tr>
<td>Marital Status</td>
<td>0 – Single</td>
</tr>
<tr>
<td></td>
<td>1 – Married</td>
</tr>
<tr>
<td>Age</td>
<td>Value from 20 to 100</td>
</tr>
<tr>
<td>Location</td>
<td>0 – Urbanized area</td>
</tr>
<tr>
<td></td>
<td>1 – Urban cluster</td>
</tr>
<tr>
<td></td>
<td>2 – Rural</td>
</tr>
</tbody>
</table>
3.2.2 Family Income

The variable for family income was used to provide information about the patient’s economic background. To calculate the family income, the income brackets for the number of earners per family from the U.S. Census Bureau were used (“Family Income,” 2016). The patient’s marital status and employment status were used to determine the number of earners per household. For this study, zero-earner households consisted of both single women who were not participating in the labor force and single women who were classified as unemployed. One-earner households were comprised of single women who were employed and married women who were either unemployed or not participating in the labor force. Two-earner households consisted only of women who were married and employed, assuming the spouse was employed as well. Patients received a randomized income value within condensed income brackets based on the distribution among the number of household earners.

3.2.3 Cancer History

The variable for history of cancer in the family was included to incorporate a level of experience with the diagnosis and treatment process. To provide this information, a study where U.S. households were surveyed to determine how many people had experienced cancer with their first-degree relatives was used (Ramsey, Yoon, Moonesinghe, & Khoury, 2006). Because the probability of this experience increases with age, the probabilities within various age groups were examined. This distribution
was translated to the WDBC dataset with patients receiving a value of 0 for those with no family history of cancer and 1 for those with a family history of cancer.

3.2.4 Preference

The variable for preference level describes the lowest level of probability at which the patient desires to be informed that the mass present is malignant. Patients were assigned a value of 0 if the patient only desired to know their probability of cancer if it is over 50%, a value of 1 if it is between 50% and 30%, a value of 2 if it is between 30% and 10%, or a value of 3 if it is under 10%. For the purpose of this study, each of the preference levels were distributed evenly among the sample.

3.2.5 Presence of Dependents

The variable for presence of dependents details if a patient has at least one dependent. Given the scope of this study and the information provided by the U.S. Census Bureau (“America’s Families and Living Arrangements,” 2017), a dependent is considered to be a child under the age of 18 living as a co-resident with the mother. These probabilities were calculated using the female percentage (50.8%) of the 2017 U.S. estimated population (“U.S. Census Bureau QuickFacts,” 2017). The probability of having a dependent varied with age, so the probabilities within separate age brackets were used. The patient received a value of 0 to indicate an absence of dependents and a value of 1 to indicate a presence of at least one dependent in the household.

3.2.6 Level of Education and Employment Status

The level of education and employment status variables were used to give insight into the patient's economic and educational background. The U.S. Census Bureau was
used to find the distribution of education level and, as a result, the distribution of employment status among the population (“Educational Attainment in the United States,” 2017). The population base for this survey was non-institutionalized civilian females over the age of 25. The U.S. Census Bureau describes non-institutionalized civilians as “U.S. civilians not residing in institutional group quarters facilities such as correctional institutions, juvenile facilities, skilled nursing facilities, and other long-term care living arrangements (“Glossary: Civilian noninstitutionalized population,” n.d.). The members of the population for this table were labeled as either not participating in the labor force, unemployed, or employed across various levels of education. The U.S. Census Bureau defines those not participating in the labor force as comprised mostly of students, housewives, and retired workers (“Glossary: Not in labor force,” n.d.). The various education levels were condensed into the following categories: not obtaining a high school diploma or GED, obtaining a diploma or GED but not attending college, attending some college or receiving an Associates degree but not a four-year degree, and receiving a four-year degree or higher (i.e. Bachelors, Masters, Ph.D., Professional). Based on these categories and proportions, the distribution of education levels among women was found. The distribution of employment status was then calculated from the condensed education levels.

3.2.7 Marital Status

The variable for marital status was used to indicate whether or not a patient is married. Because the likelihood of being married varies with age, the probabilities within various age brackets among females in 2017 from the U.S. Census Bureau was used (“Table A1. Marital Status Of People 15 Years And Over, By Age, Sex, and Personal
A woman was considered single and assigned a value of 0 to indicate she was widowed, divorced, separated, or had never been married. She was considered married and received a value of 1 in instances where she was legally married, whether the spouse was present or absent.

3.2.8 Age

The variable age was included because the probability of being diagnosed with breast cancer increases immensely with age. Using the 2014 data from the North American Association of Central Cancer Registries incidence data ("NAACCR Fast Stats 2010-2014 Cancer Incidence Data," 2018), the probability within various age brackets was used to assign patient ages. The randomly generated ages based on these various probabilities fell between 20 years and 100 years.

3.2.9 Location

The variable for geographical location was used to provide information about the patient’s environment. Values from the U.S. Census Bureau’s Urban, Urbanized Area, Urban Cluster, and Rural Population, 2010 and 2000: United States were used to create this variable ("Urban, Urbanized Area, Urban Cluster, and Rural Population, 2010 and 2000: United States," 2015). The 2010 U.S. Census Bureau considers an urbanized area as one containing 50,000 or more people, an urban cluster as an area containing more than 2,500 and less than 50,000 people, and rural area as one containing fewer than 2,500 people. Each patient was assigned either a value of 0 for living in an urbanized area, a value of 1 for living in an urban cluster, or a value of 2 for living in rural area.
3.3 Randomized Data Generation

DataWorks’ MATLAB was used to randomize the nine additional attributes, and a new empty dataset was created. To create the randomly generated dataset, an empty set was created, and the values were generated in a loop. A random number (RN) between 0 and 1 was generated. To produce the age, if statements specifying the probability that a patient should fall within a given range were created. When the RN found the appropriate probability range, a random age within the given range was produced. Like with the age attribute, to produce the categorical attributes that depended only on probability—location and education—a new RN was created and assigned to the correct probability range, resulting in the appropriate categorical value. The remaining attributes were dependent either on age or some other attribute. Conditional if statements were again constructed; when all RN and other attribute conditions were met, their categorical values were assigned (or randomized within a range as with income). This process was repeated in the loop for all 56,900 patients. Once all the attributes were created, they were joined with the 11 attributes from the WDBC dataset. The code for generating these non-medical attribute values can be found in Appendix A.

3.4 SOM Construction

To begin SOM construction, the new data containing both the WDBC measures and the generated attributes was loaded and initialized. Using the ‘randperm’ function, the order of patients was randomized. Each column was then transformed to be a value between -1 and 1, so all data points would be on the same scale. The dataset was separated into the classification column (benign or malignant) and the rest of the attributes to be used as input for the SOM, excluding the patient identification number.
Additionally, it was split into a training set containing 56,880 patients and a test set containing the remaining 20 patients.

The SOM was initialized by setting the size to a $d \times d$ grid with ‘dimensions,’ the number of training steps for the initial covering of the input space with ‘coverSteps,’ the size of the initial neighborhood with ‘initNeighbor,’ the layer topology with ‘topologyFcn,’ and the method used to calculate the distance with ‘distanceFcn.’ The SOM net was initialized with the ‘selforgmap’ function. The ‘train’ function was then executed to expose each patient to the SOM, resulting in each patient being exposed 200 times. The initialization values were adjusted, and the output was studied to determine the optimum values as follows: ‘dimensions’ set to 12x12, ‘coverSteps’ set to 30, ‘initNeighbor’ set to 3, ‘topologyFcn’ set to a square grid top, and ‘distanceFcn’ set to Euclidean. Additionally, a weight of 5.5 was given to the mass measurements from the original dataset and the additional generated attributes were left with a weight of 1; this was done to ensure that most of the classification was determined by the medical attributes. After finalizing these inputs, several plots were generated with ‘plotsomhits’ and ‘plotsomplanes.’ A new structure was also created using instances so each patient within a cluster could be viewed individually and the cluster could be profiled. The code for generating the SOM can be found in Appendix B.
CHAPTER IV
ANALYSIS

Once the model was created, it was used to visualize the maps in various contexts.

Figure 4.1 below shows the hit map containing the 144 clusters that were created and the number of patients assigned to each cluster.

Figure 4.1 SOM train hits map
A set of plane subplots was generated—one for each attribute—displaying the weights from each input to the output layer’s neurons. Figure 4.2 shows these plots for the breast mass measurements from the WDBC database.

![Figure 4.2 Weight plots for WDBC medical attributes](image)

Most of these measurement attributes pinpoint the same regions of the map as having similar traits. Similar patterns can be seen among inputs 1, 3, 4, 6, 7, and 8, indicating that these measurement attributes are most responsible for creating the SOM classifications. Figure 4.3 shows the same kind of plots applied to the non-medical generated attributes.
It is important to again note the random generation of this data plays a role in its somewhat scattered appearance. The plots in Figure 4.3 show age, location, preference level, marital status, education level, employment status, presence of dependents, history of cancer, and income weights, respectively. The plots for age (input 11), marital status (input 14), education level (input 15), employment status (input 16), history of cancer (input 18), and income (input 19) also pinpoint the same regions of the map as having similar characteristics.

After analyzing the weight plots for each attribute, a new set of plots was produced to evaluate the 144 clusters individually. For each of the additional randomly generated attributes, a map was created that showed the average value of all hits/patients within each cluster on the map. The most useful of these maps was the mean map for the classification label, shown in Figure 4.4.
This map displays the average label value for each cluster; white clusters with a value of -1.00 indicate that every patient in that cluster possessed a benign tumor, while black clusters with a value of 1.00 indicate that every patient in that cluster possessed a malignant tumor. Every other cluster shows that either most in that cluster were benign, most were malignant, or there was an even mix. This same method of producing a map displaying the average value for each cluster was applied to each attribute. The remaining maps can be found in Appendix C.
CHAPTER V
APPLICATIONS

To evaluate the applications of this tool, the test set containing twenty entries was run through the SOM to simulate how decisions can be made when new patients are evaluated based on the SOM framework. The hit map for the test dataset containing twenty entries is seen in Figure 5.1.

![Test Set Hit Map](image)

Figure 5.1 Test set hit map

The test set hit map was then superimposed with the mean label map from Figure 4.4 to show which clusters the twenty “new” patients fell into. This superimposed map can be seen in Figure 5.2.
Figure 5.2 Superimposed test set hit map and mean label map

This superimposed map details the average diagnosis for each cluster with green circles placed around the hit map clusters. Clearly, the hits in the black clusters will likely be diagnosed as malignant and the hits in the white clusters will likely be diagnosed as benign. Table 5.2 details the cluster for each of the 20 “new” patients, as well as the mean class label for each cluster and the actual class label for each patient.
Table 5.1  Cluster class evaluation

<table>
<thead>
<tr>
<th>Index</th>
<th>Cluster #</th>
<th>Mean Cluster Class</th>
<th>Actual Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>112</td>
<td>-0.95</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>52</td>
<td>-0.91</td>
<td>-1</td>
</tr>
<tr>
<td>4</td>
<td>86</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>7</td>
<td>66</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>114</td>
<td>-0.85</td>
<td>-1</td>
</tr>
<tr>
<td>9</td>
<td>112</td>
<td>-0.95</td>
<td>-1</td>
</tr>
<tr>
<td>10</td>
<td>137</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>11</td>
<td>111</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>0.17</td>
<td>-1</td>
</tr>
<tr>
<td>13</td>
<td>114</td>
<td>-0.85</td>
<td>-1</td>
</tr>
<tr>
<td>14</td>
<td>139</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>15</td>
<td>92</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>131</td>
<td>0.46</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>68</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>48</td>
<td>-0.33</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>41</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>77</td>
<td>0.45</td>
<td>1</td>
</tr>
</tbody>
</table>

For evaluation purposes, if a prediction was within 10% of the actual class label, it was counted as accurate. This method results in an accuracy of 80%. However, the decision-making focus of this research draws attention to the bordering hits in clusters that are not all benign or malignant. For demonstrative purposes, this analysis focuses on three of these patient hits, located in cluster number 3 with a label mean of 0.17, cluster number 52 with a label mean of -0.91, and cluster number 131 with a label mean of 0.46.
The means for each of the non-categorical non-medical attributes for these three clusters were extracted and the clusters were profiled. Table 5.3 displays the averages for each attribute in each cluster.

Table 5.2 Test patient cluster means for non-categorical attributes

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Label</th>
<th>Age</th>
<th>Family Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.17</td>
<td>71.15</td>
<td>$93,385.60</td>
</tr>
<tr>
<td>52</td>
<td>-0.91</td>
<td>73.84</td>
<td>$61,896.94</td>
</tr>
<tr>
<td>131</td>
<td>0.46</td>
<td>66.63</td>
<td>$128,267.70</td>
</tr>
</tbody>
</table>

The first major difference seen among the three clusters is the label value. The patient placed in cluster 52 is very likely to have a benign tumor, while the patients in clusters 3 and 131 are more likely to possess malignant tumors. While this is valuable information, the decision-making focus is more on the non-medical attribute differences among the clusters. The average age of patients in clusters 3 and 52 is higher than that of patients in cluster 131. Finally, the means for family income differ significantly across each cluster.

To display the differences among clusters for each categorical attribute, percent mixes were created. Table 5.4 displays these mixes.
By combining the information from tables 5.3 and 5.4, the following profiles for each cluster can be generated. There is little distinction among the mixes in each of the clusters for the preference, location, and dependents attributes. All three clusters have an even mix for preference level with each value making up around 25% of the cluster. Similarly, around 70% of patients in each of the clusters live in urbanized areas, 10% living in urban clusters, and 20% live in rural areas. The education level is also similarly distributed, with around 40% possessing a Bachelors degree or higher, 25% possessing only some college experience, 25% having only a high school diploma (or equivalent),
and 10% having not completed high school. Finally, around 99% of patients in each of the three clusters do not have dependents.

Patients in cluster 3 have an average age of 71 and around 40% have a family history of cancer. There is a relatively even mix of those who are married versus single, though there is a slightly higher number of married individuals at 58%. A little less than 70% of these patients are employed, with 30% not participating in the labor force and 2% unemployed. As a result, the average family income for this cluster is between $90,000 and $95,000.

The average age of patients in cluster 52 is around 74. Only 15% of patients in this cluster have a history of cancer in the family. All patients in this cluster are unmarried. Similar to cluster 3, about 70% of them are employed, with 25% not participating in the labor force and only 3% unemployed. As a result, the average income in this cluster is lower at around $62,000.

The average age of patients in cluster 131 is lower than the other two clusters at 67. Similar to cluster 33% have a family history of cancer. Unlike cluster 52, all patients in this cluster are married. Among these patients, 97% are employed, 3% are unemployed, and all are labor force participants. As a result, the average family income in this cluster is much higher at around $130,000.

Given these three patient profiles, the diagnosis and treatment plan would be decided and delivered differently. The patient in cluster 3 that has an average income of $93,000 may not be able to afford the same treatment options as the patient in cluster 131 with an average income almost $40,000 higher. Additionally, less than half of the cluster
3 patients have a spouse to serve as a support system, while everyone in cluster 131 has that support system.

Combining this framework with real data would create a powerful decision-making tool. The differences in the attributes among the three clusters selected for evaluation are not present for every attribute. The inclusion of real data would only improve the decision-making capabilities. For instance, a difference among preference levels would aid in determining whether or not to inform the patient if they belonged to a cluster with a malignancy probability of 15%. Similarly, a cluster with a high rate for presence of dependents may be offered different treatment options than those with no dependents. Those with no family history of cancer may not understand the implications of different treatment options and may have a higher anxiety level, resulting in a different delivery method. Discrepancies between categories for employment status could affect what level of treatment someone is able to receive due to time availability.

These conclusions along with many others can be made and used to tailor the diagnosis, delivery, and treatment options to each individual patient. This allows for a more personalized experience for patients, increasing their satisfaction and the trust between patients and their doctors. This aspect of precision medicine can also increase both doctor and patient confidence that the best treatment option is chosen.
CHAPTER VI
LIMITATIONS AND FURTHER RESEARCH

The framework nature of this study provides a natural next step for further research. The largest limitation in this study was the lack of availability of real data, resulting in random data generation based on population statistics. The data was generated to resemble the population as closely as possible, but it is not a sufficient substitution for real patient data. Using this framework, future studies can focus on the collection of the nine non-medical attributes detailed in this study.

Additionally, other attributes beyond the nine developed in this study should be considered. One aspect of the patient profiles that was unable to be taken into consideration is personality. While attributes pertaining to patients’ lifestyles were included in the study and do aid the physician in developing a treatment plan, the profiles can be further tailored. A study to focus on the quantification of personality type, including level of resiliency and anxiety, could provide additional personalization of the diagnosis and treatment plans.

A final clear avenue for further research would be the incorporation of physicians. After including real data and any other attributes, in addition to personalities, most of the population will be represented on the map. This will allow for doctors to suggest diagnosis deliveries and treatment plans for each cluster based on the full patient profiles.
REFERENCES


31
APPENDIX A

NON-MEDICAL DATA GENERATION CODE
The following figures (A.1, A.2, A.3, A.4, A.5, A.6, and A.7) detail the code used to generate the age, preference, location, marital status, education level, employment status, presence of dependents, cancer history, and income attributes based on population statistics.

```matlab
NewData = [];
for i=1:369
  for j=1:100
    firstsegment = OldData(i,:);
    %Randomly creating age
    rn.rand();
    if(rn<.07)
      Age=20+randi(29);
    end
    if(rn>.07 & rn<.345)
      Age=50+randi(14);
    end
    if(rn>.345 & rn<.61)
      Age=65+randi(8);
    end
    if(rn>.61)
      Age=75+randi(25);
    end
    %Randomly creating preference
    Preference=randi(4)-1;
    %Randomly creating location
    rn=rand();
    if(rn<.71)
      Location=0;
    end
    if(rn>.71 & rn<.81)
      Location=1;
    end
    if(rn>.81)
      Location = 2;
    end

Figure A.1  Age, preference, and location generation
```
Figure A.2  Marital status and education level generation

```plaintext
# Randomly creating marital status

m <- rand();
if(Age>=18 && Age<30)
  if(m<=.256)
    Marital=1;
  end
if(m>.256)
  Marital=0;
end
if(Age>=30 && Age<40)
  if(m<=.597)
    Marital=1;
  end
if(m>.597)
  Marital=0;
end
if(Age>=40 && Age<75)
  if(m<=.631)
    Marital=1;
  end
if(m>.631)
  Marital=0;
end
if(Age>=75 && Age<101)
  if(m<=.408)
    Marital=1;
  end
if(m>.408)
  Marital=0;
end
if(Age>=101)
  if(m<=.152)
    Marital=1;
  end
if(m>.152)
  Marital=0;
end

# Randomly creating education

n <- rand();
if(n<=1)
  Education=0;
end
if(n>.1 && n<=.377)
  Education=1;
end
if(n>.377 && n<=.654)
  Education=2;
end
if(n>.654)
  Education=3;
end
```
%Randomly creating employment

rn=rand();
if(Education==0)
  if(rn<.665)
    Employment=0;
  end
  if(rn>.665 && rn<.693)
    Employment=1;
  end
  if(rn>.693)
    Employment=2;
  end
end

if(Education==1)
  if(rn<.519)
    Employment=0;
  end
  if(rn>.519 && rn<.545)
    Employment=1;
  end
  if(rn>.545)
    Employment=2;
  end
end

if(Education==2)
  if(rn<.386)
    Employment=0;
  end
  if(rn>.386 && rn<.408)
    Employment=1;
  end
  if(rn>.408)
    Employment=2;
  end
end

if(Education==3)
  if(rn<.296)
    Employment=0;
  end
  if(rn>.296 && rn<.313)
    Employment=1;
  end
  if(rn>.313)
    Employment=2;
  end
end

Figure A.3   Employment status generation
%Randomly creating dependents
rn=rand();
if(Age<25)
  if(rn<.017)
    Dependents=1;
  end
if(rn>.017)
  Dependents=0;
end
if(Age>=24 && Age<30)
  if(rn<.041)
    Dependents=1;
  end
if(rn>.041)
  Dependents=0;
end
if(Age>=29 && Age<35)
  if(rn<.065)
    Dependents=1;
  end
if(rn>.065)
  Dependents=0;
end
if(Age>=34 && Age<40)
  if(rn<.079)
    Dependents=1;
  end
if(rn>.079)
  Dependents=0;
end
if(Age>=39 && Age<45)
  if(rn<.073)
    Dependents=1;
  end
if(rn>.073)
  Dependents=0;
end
if(Age>=44 && Age<50)
  if(rn<.059)
    Dependents=1;
  end
if(rn>.059)
  Dependents=0;
end

Figure A.4  Presence of dependents generation
Figure A.5  Presence of dependents (cont.) and cancer history generation
Figure A.6  Cancer history generation (cont.)

```plaintext
258 - if (Age>49 && Age<60)
259 -   if (rn<.324)
260 -     History=1;
261 -   end
262 -   if (rn>.324)
263 -     History=0;
264 - end
265 - end
266 - if (Age>59 && Age<70)
267 -   if (rn<.344)
268 -     History=1;
269 - end
270 -   if (rn>.344)
271 -     History=0;
272 - end
273 - end
274 - if (Age>69)
275 -   if (rn<.382)
276 -     History=1;
277 - end
278 -   if (rn>.382)
279 -     History=0;
280 - end
281 - end
282`
```
Randomly creating income

```c
# Randomly creating income

    if (Marital==0 & Employment==0) || ...
        if (Marital==0 & Employment==1)
            if (rnc<.654)
                Income=0+randi(49999);  
                end
            if (rnc>.654 & rnc<.907)
                Income=50000+randi(459999);
                end
            if (rnc>.907 & rnc<.987)
                Income=100000+randi(99999);
                end
            if (rnc>.987)
                Income=200000+randi(16666669);
                end
        end
    if (Marital==0 & Employment==2) || (Marital==1 & ...
        if (Marital==0 & Employment==0) || (Marital==1 & Employment==1)
            if (rnc<.486)
                Income=0+randi(49999);
                end
            if (rnc>.486 & rnc<.799)
                Income=50000+randi(459999);
                end
            if (rnc>.799 & rnc<.947)
                Income=100000+randi(9999999);
                end
            if (rnc>.947)
                Income=200000+randi(16666669);
                end
        end
    if (Marital==1 & Employment==2)
        if (rnc<.15)
            Income=0+randi(49999);
            end
        if (rnc>.15 & rnc<.515)
            Income=50000+randi(999999);
            end
        if (rnc>.515 & rnc<.869)
            Income=100000+randi(16666669);
            end
        if (rnc>.869)
            Income=200000+randi(16666669);
            end
    end

        secondsegment=[Age, Preference, Location, Marital, ...
                      Education, Employment, Dependents, History, Income];
        NewRow=[firstsegment, secondsegment];
        NewData=[NewData; NewRow];
        end
```

Figure A.7  Family income generation
APPENDIX B

SOM CODE
The following figures (B.1, B.2, B.3, B.4, and B.5) detail the code used to initialize, train, and map the SOM.

```matlab
load('C:\Users\vvs14.ISE-BCY7JH2\OneDrive\MSU 2017-2018\Thesis\SOM\Vanga_Code\GeneratedData.mat');
Data = NewData;

% Randomize the order of the dataset
RandPERM = randperm(56900);
Data = Data(RandPERM,:);

Class = Data(:,2);
InData = Data(:,3:21);

Age = InData(:,11);
Preference = InData(:,12);
Location = InData(:,13);
MaritalStatus = InData(:,14);
Education = InData(:,15);
Employment = InData(:,16);
Dependent = InData(:,17);
History = InData(:,18);
Income = InData(:,19);

% Transform every value to be between -1 and 1
NewData = [];
for i=1:19
    Container = InData(:,i);
    Container = Container - min(Container);
    Container = (Container - max(Container)/2)/max(Container)*2;
    NewData(:,i) = Container;
end

InData = NewData;
InData = [5.5*InData(:,1:10), InData(:,11:19)];

ClassTrain = Class(1:56880,:);
%ClassTest = Class(56881:end,:);

AttributeTrain = InData(1:56880,:);
%AttributeTest = InData(56881:end,:);
```

Figure B.1 Randomizing order, assigning attributes, transforming values, and creating train set
Figure B.2 Initializing, training, and plotting SOM, and preparing heat map
empty_Cluster.Index = [];
Clusters = repmat(empty_Cluster, [1,D^2]);

for i=1:D^2
    Clusters(i).Index = i;
    WMatrix = [];
    WStruct = Instances([Instances.Cluster]==i);
    Clusters(i).Label = mean([WStruct.Label]);
    Clusters(i).Age = mean([WStruct.Age]);
    Clusters(i).Preference = mean([WStruct.Preference]);
    Clusters(i).Location = mean([WStruct.Location]);
    Clusters(i).MaritalStatus = mean([WStruct.MaritalStatus]);
    Clusters(i).Education = mean([WStruct.Education]);
    Clusters(i).Employment = mean([WStruct.Employment]);
    Clusters(i).Dependents = mean([WStruct.Dependents]);
    Clusters(i).History = mean([WStruct.History]);
    Clusters(i).Income = mean([WStruct.Income]);
end

%Label%
Label_M = zeros(D,D);
Age_M = zeros(D,D);
Preference_M = zeros(D,D);
Location_M = zeros(D,D);
MaritalStatus_M = zeros(D,D);
Education_M = zeros(D,D);
Employment_M = zeros(D,D);
Dependents_M = zeros(D,D);
History_M = zeros(D,D);
Income_M = zeros(D,D);
for i=1:D^2
    WV = i;
    yi = floor(WV/D);
    xi = WV-yi*D;
    yi = yi+1;
    if(xi==0)
        xi = D;
        yi = yi-1;
    end

    Label_M(D+1-yi,x1) = Clusters(i).Label;
    Age_M(D+1-yi,x1) = Clusters(i).Age;
    Preference_M(D+1-yi,x1) = Clusters(i).Preference;
    Location_M(D+1-yi,x1) = Clusters(i).Location;
    MaritalStatus_M(D+1-yi,x1) = Clusters(i).MaritalStatus;
    Education_M(D+1-yi,x1) = Clusters(i).Education;
    Employment_M(D+1-yi,x1) = Clusters(i).Employment;
    Dependents_M(D+1-yi,x1) = Clusters(i).Dependents;
    History_M(D+1-yi,x1) = Clusters(i).History;
    Income_M(D+1-yi,x1) = Clusters(i).Income;
end

close all;
Figure B.5    Creating heat map and plotting test data

```matlab
mat = Label_N;
mat = mat;
x = [1 D];
y = [1 D];

im = im(x,y,mat); %Create a colored plot of the matrix values
colormap(flipud(hot)); %Change the colormap to gray (so higher values are
%black and lower values are white)
colorbar;

textStrings = numstr(mat(:,[:,.2F])); %Create strings from matrix values
textStrings = strtrim(cellstr(textStrings)); %Remove any space padding
[x,y] = meshgrid(1:D); %Create x and y coordinates for the strings
hStrings = text(x(:,y(:,textStrings(:,... %Plot the strings
'HorizontalAlignment','center');
midValue = mean(get(gca,'CLim')); %Get the middle value of color range
textBoxColor = repmat(mat(:, > midValue,1,3)); %Choose white or black for
%text color of the strings so
%they can be easily seen over
%the background color
set(hStrings,'Color',num2cell(textBoxColor,2)); %Change the text colors
set(gca,'XTick',1:7,... %Change the axes tick marks and tick labels
'XTickLabel',{"","","",""},...
'YTick',1:5,...
'YTickLabel',{"","","",""},...
'TickLength',[0 0]);

%Test set

plotcorners(net,AttributeTest');

y = net(AttributeTest');
Cluster_numbers = vec2ind(y');
```
APPENDIX C

NON-MEDICAL ATTRIBUTE MEAN MAPS
The following figures (C.1, C.2, C.3, C.4, C.5, C.6, C.7, C.8, and C.9) detail the mean maps for each of the nine generated attributes. The code used to generate each of these maps is found in line 161 of Figure B.5. In each case, the word ‘Label’ was replaced with the attribute name.

Figure C.1  Mean map for age
Figure C.2  Mean map for location

Figure C.3  Mean map for preference level
Figure C.4  Mean map for marital status

Figure C.5  Mean map for education level
Figure C.6  Mean map for employment status

Figure C.7  Mean map for presence of dependents
Figure C.8 Mean map for history of cancer

Mean map for family income