Differential Predictive Modeling for Racial Disparities in Breast Cancer

Indranil Palit[1], Chandan K. Reddy[1] and Kendra L. Schwartz[2]
[1]*Department of Computer Science, Wayne State University, Detroit, MI 48202, USA.*
[2]*Karmanos Cancer Institute, Detroit, MI 48201, USA.*
E-mail: indranilpalit@wayne.edu, reddy@cs.wayne.edu, kensch@med.wayne.edu

Abstract

The primary objective of disparities research is to model the differences across multiple groups and identify the groups that behave significantly different from each other. Independently generating various decision trees for different subsets of the data will not allow us to study the impact of the various attributes on these different subgroups. We propose a novel technique for inducing similar decision trees for different subpopulations and also develop a new distance metric between two decision trees which measures the difference in the underlying data distributions of these subgroups. The proposed framework is evaluated by analyzing the racial disparities in breast cancer. Our method was able to rank different populations with respect to the disparity and detect the attributes that are most responsible for such differences.

1. Introduction

According to National Cancer Institute, health disparities are differences in the incidence, prevalence, mortality, and burden of cancer and related adverse health conditions that exist among specific population groups in the United States [8]. These population groups may be characterized by gender, age, ethnicity, education, income, social class, disability, geographic location, or sexual orientation. There are many factors that influence disparity. In the context of racial differences in the United States, it is found that African Americans, Native Americans, and Latinos have higher incidence of chronic diseases, higher mortality, and poorer health outcomes compared to the white population [9]. Cancer incidence rate among African Americans (AA) is 10% higher than Caucasian Americans (CA) [10]. It is also reported that AAs and Latinos have approximately twice the risk compared to the CAs for developing diabetes. Minorities also have higher rates of cardiovascular disease, HIV/AIDS, and infant mortality than CAs [10]. There has also been some research work in locating disparity among populations from different geographic locations. Some works describe the overall cancer incidence, mortality, and prevalence, age-adjusted temporal trends, and age-specific incidence patterns in selected geographic locations [11]. Disparity analysis on different demographic and socio economic factors such as race, socio-economic position, sexual practice, age, access to clinical trials, literacy, disability, immigration status, insurance, and geography have been studied [12]. Most of the disparity analyses are centered on socio economic factors. A recent study of 20,000 cancer patients [13] in the United States found that AAs compared to the European Americans are less likely to survive breast, prostate and ovarian cancers even when equal care is given. This suggests that other biological factors may be important for this racial disparity. One of main goals of disparities research is to model the differences across multiple groups of patients and identify those populations where these groups behave differently.

In this paper, we will build novel predictive modeling tools to capture racial disparities in breast cancer. Breast cancer is the most common cancer and the second leading cause of cancer death among US women [14]. For several decades, researchers have documented disparities across several domains in breast cancer risk, incidence, screening, diagnosis, treatment, survival and mortality. There is tremendous public support, especially among women, for increasing efforts to find a cure for the breast cancer. However, these efforts are unlikely to play a significant role in addressing disparities in risk, treatment, and mortality unless there is a better understanding and recognition of specific factors leading to disparities.

We develop a new method for modeling differences across subgroups. In this work, we used a decision tree model to measure dissimilarity between groups of data, which in turn are portions of a larger dataset. The traditional solutions in the field of disparities research partially addresses the dissimilarity issue, but fails to provide any comprehensive technique in terms of the prediction models. They build basic statistics of the

data that belongs to each group to show the differences in the distribution. It is vital to develop an integrated framework that can model the disparity and simultaneously develop a predictive model that will not only provide the statistics of the differences but also give more insights into the attributes that are precisely bringing the difference in terms of the predictive model. Hence, we build *'differential predictive models'* that can simultaneously capture and model the differences amongst different subgroups in the data and their corresponding predictive models.

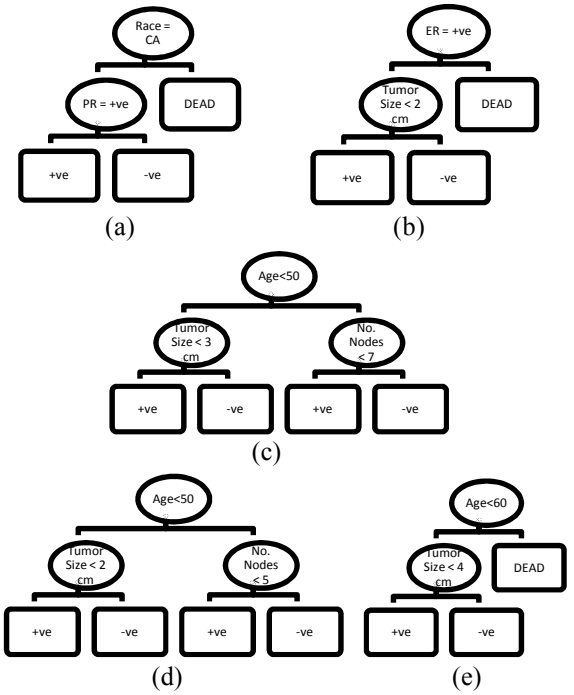

Fig 1: An example to illustrate the need for the proposed work.

To achieve this goal, we used a decision tree model which is a very well known data classification tool. Decision trees not only induce accurate classification models, but also provide a good representation of the data itself [1]. Compared to other data mining models such as support vector machines or neural networks, the decision trees can yield an insightful model for the analysis that is readily interpretable by the domain experts. We devise new measures to compute the group differences for each population and then quantitatively compare those differences. Group differences are measured by observing the difference between the Decision trees induced for different groups.

Decision trees are multivariate representation of the data [2]. Data can be represented by many tree models. However, not all of these models perform well in terms of the prediction accuracy. Traditional tree induction algorithms search for the most optimal model from the model space. Typically, there exist several tree models in the model space that are optimal or very close to the optimal model in terms of a particular evaluation metric such as accuracy. It is possible to find some other representations for the data other than the induced model. In this work, we demonstrate a new technique to generate such alternative models. The models that are generated will be structurally different, yet will be a good representation of the data as long as the predictive accuracy is within an acceptable margin. Using our technique, we can find two such trees for the two groups which are structurally as similar as possible. Then, the distance between these two trees is measured and used as a metric of difference between the underlying datasets. Using this approach, we will be able to measure the group difference in a particular region and compare these differences across other regions. *This is useful in ranking the regions where the disparity values are significantly different.*

Existing tree induction methods do not ensure that the trees induced from different subsets of the data can be comparably similar even if the data distribution is almost same. The purpose of the decision tree is to build a predictive model as accurate as possible by maximizing the predictive accuracy at each level of the tree induction process. Decision trees are very sensitive to minor changes in the dataset [3] and hence are regarded as 'unstable classifiers'. Even for datasets with similar distribution, the induced trees can be structurally dissimilar due to either noise or sensitivity of the split measures to a slight difference in the attribute values. Measuring the distance between these trees and interpreting it as a distance metric for the underlying data distribution will be misleading. Hence, it is important to induce tree models that are structurally close to each other without significantly compromising on the accuracy of the prediction. *Fig. 1* illustrates an example. *Fig. 1(a,b)* show the two decision trees induced independently and hence will have no structural similarity. The distance between these two trees will be high even if the data distributions are the same. *Fig. 1(c)* shows the tree induced from complete datasets and is regarded as the reference tree. *Fig. 1(d,e)* show the trees for the two groups induced by our algorithm based on the reference tree. Now, the two trees are structurally similar and their distance can be used a quantitative metric to measure the differences in the subpopulations. We propose a new transition model for these decision trees to obtain a continuum of tree models which can then be used to make comparisons across different subgroups.

2. Differential Predictive Modeling

The C4.5 decision tree induction algorithm [1] constructs a decision tree from a given training set. It is considered to be one of the popular data mining algorithms [4]. The standard performance measure of a decision tree over a set of instances is the classification error which is the percentage of misclassified instances. As mentioned earlier, C4.5 does not enforce similar structure between the models for different groups. Generating structurally similar trees for the two different subgroups is a very important building block of our approach. To ensure the similarity of the structure of these trees, we construct a reference decision tree that is induced from the complete data set consisting of instances from both the groups of interest.

2.1. Similar Tree Induction Method

We will first introduce some basic notation needed to comprehend our algorithm. Let us denote these two groups as $group_1$ and $group_2$. Let D_1 and D_2 contain the instances of $group_1$ and $group_2$ respectively. Both D_1 and D_2 have the same attribute space $\grave{A} = \{A_1, \ldots. A_n\}$ where *n* is the number of attributes. For D_1 and D_2, T_1' and T_2' are the C4.5 decision trees; T_1 and T_2 are the similar decision trees that will be induced by our approach. Let T be the tree induced from the data set $D_1 \cup D_2$ using the C4.5 algorithm. We call it a reference tree. T is a good representation for both D_1 and D_2, as the data set upon which T is trained includes both D_1 and D_2. So, we assume that, a tree model which is structurally similar to T, will also reasonably explain D_1 in terms of prediction accuracy (shown in the results section). So, in our approach we generate T_1 (for $group_1$) and T_2 (for $group_2$), both of which try to mimic the structure of T as much as possible.

Standard C4.5 selects the attribute with the highest information gain ratio at a given node in the tree during the induction. Instead of using information gain ratio as the split criteria, we introduce another metric, total Weight ($\grave{W}$) as the modified split criteria. The total Weight ($\grave{W}$) is a weighted summation of the information gain ratio and a similarity score ($\acute{S}$), where $i \in [1, n]$:

$$\grave{W}_i = w \times GainRatio_i + (1 - w) \times \acute{S}_i$$

For each of the *n* attributes, both the Gain Ratio and Similarity score are normalized between 0 and 1 using Min-Max normalization [5]. The 'w' is a weight parameter which varies from 0 to 1. When w = 1, the same tree as induced by C4.5 (T_1' or T_2') is obtained. If the value of w is not equal to 1, then it will result in a structurally different tree than T_1' or T_2'. It should be noted that the trees thus obtained, would be more similar to the reference tree T.

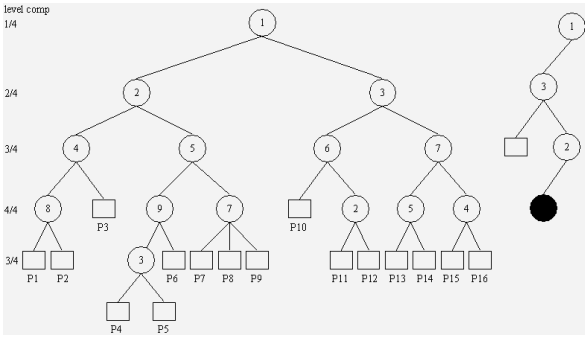

Fig 2: An example reference tree structure, T and the tree T_1 expanded so far.

2.2. Similarity Score for an attribute

As explained before, we need a similarity score for each attribute to decide upon the potential split attribute. The similarity score proposed in this work has two components: (i) Level component and (ii) Path component.

2.2.1. Level component of the similarity score. While inducing the trees for the groups (T_1 or T_2), our algorithm prefers the attributes appearing at the same level of the reference tree (T). To reflect this preference in the algorithm, we assign the similarity score to all the attributes in À, which is proportional to their relative depth in T. For example, if we are considering a node in T_1 for expansion whose depth in T_1 is d; then a maximum similarity score (which is 1) is assigned to the attributes appearing in the same depth (d) in the reference tree (T). The Attributes appearing in depth (d+1) and (d-1) will be assigned a similarity score which is less than the similarity score for depth d but greater than that for depths (d+2) or (d-2). Thus, all the attributes appearing in T will be assigned some non-zero positive score based on their position. Attributes in À that do not appear anywhere in T will be scored zero. If the same attribute appears at a different depth in the decision tree T then it will be assigned a score only once and this score is given to its nearest position. As shown in *Fig 2,* the tree T_1 is being expanded at depth 3(solid node). So, in the reference tree T; the attributes 2, 4, 5, 7, 8, 9 (depth 3) get a score of 1. Attributes 3 (depth 4) and 6 (depth 2) are assigned a score of 0.75. Other attributes of depth 2 have already been assigned a higher score at depth 3. Finally, attribute 1 gets a score of 0.25 (depth 0).

2.2.2. Path component of the similarity score. We also assign a similarity score to each attribute by performing path-wise analysis. Since the C4.5 performs a Depth First Search (DFS) to construct the tree, it explores a path fully down to the leaf node and then backtracks to an earlier node which is not fully expanded, thus adding a new path. Our algorithm also proceeds in the same manner.

While constructing T_1 using our algorithm, let us consider a path þ. The objective here is to make paths of T_1 as similar as possible to the paths of T. Hence, higher scores are assigned to the attributes that would help to form similar paths. To ensure this, our algorithm compares the current path þ of T_1 with every path in the reference tree T. Attributes in those paths (other than the attributes already in þ) are assigned a similarity score proportional to the similarity of that path with þ; which is measured by the number of attribute-outcome pairs common in both the paths. If every attribute-outcome pair of þ appears in a path of T then the assigned score for that path is the maximum (which is 1). Attributes may appear in different paths of T, yielding different scores for a single attribute. In such cases, the maximum of these scores is assigned to that attribute. If þ is empty (at the very beginning when we start with the root) then no score for path component is assigned. Tree T (*Fig 2*) has 16 paths and the current path in T_1 (þ) has 3 attributes (1, 2, 3). P4, P5, P11, P12 are four different paths in T having all the three attributes. So, the other attributes of these four paths: 5, 6, 9 get a score of 1. In P1, P2 there are 2 attributes (1, 2) common with þ. So, attributes 4 and 8 are scored 2/3. For paths P7, P8, P9, the attribute 7 is given a score of 2/3. Attribute 5 has already been weighted with a higher score. Other attributes are scored in the same manner.

After the level and path components of the similarity score are computed, these two components are added for every attribute. This aggregated score is then normalized to get Ś.

2.3. A Continuum of Tree Models

First, we will discuss our algorithm for measuring the tree distance for any two decision trees T_1 and T_2. Then, we will show the use of this distance measure in generating a continuum of tree models from the reference tree.

For measuring the tree distance, we compare the attributes at the same level of two trees T_1 and T_2. At each level, both the trees contribute a distance parameter that adds up to the overall distance. The distance parameter for the tree T_1 is the ratio of the number of attributes of T_1 that are not common in the same level of T_2, to the number of attributes of T_1 at that same level. There is a similar distance parameter for T_2. If a tree is deeper than the other tree, then only the deeper tree contributes for the distance parameter at the deeper level of that tree, where there are no attributes at the same level in the other tree. The overall distance is averaged by the total number of distance contributions made by both the trees. Each of these contributions has a maximum value of 1. So, the average distance will also have a maximum value of 1. If the trees are exactly the same, then the distance will be zero.

For generating a continuum of similar models, we will use our algorithm in the following manner: Let us consider the problem of measuring the disparity of two groups. The dataset corresponding to these two groups are D_1 and D_2 respectively. Let T be the reference tree built on the dataset $D_1 \cup D_2$. For a fixed w, we would build a tree for D_1. In the process of the tree generation, we calculate the level and the path components of the similarity score repeatedly in order to determine the attribute for splitting. At the end of the tree generation process, we measure its distance from the reference tree T. Thus, for different w values, we will get a continuum of tree models for D_1. Among these models, a single model whose distance from T is the minimum will be chosen. Similarly, we build a model of continuum for D_2 and select the optimal model. Finally, the distance between these two optimal models is obtained by calculating their distance directly from each other. This distance is a measure for the difference between D_1 and D_2.

3. Experimental Results

The dataset is assembled retroactively for this study and contains data records from over 34,000 women diagnosed with primary invasive breast cancer in the Detroit area during the years 1995-2006. All of the patients in the analysis were residents of the tri-county Detroit metropolitan area (Wayne, Oakland and Macomb counties) at the time of diagnosis. The tumor behavior of these women is either in situ or malignant. The women are diagnosed with a first primary invasive breast cancer identified through the population based Metropolitan Detroit Cancer Surveillance System (MDCSS). MDCSS is one of founding members of the NCI Surveillance Epidemiology and End Results (SEER) program. We used the socio-economic status (SES) data at the census tract level to improve the precision of the SES estimates [6]. The strength of such large data sets such as the SEER database however is in the large number of cases and the long follow up available [7].

3.1. Data Preprocessing

In our dataset, the patients who had died from breast cancer during the years 1995 to 2006 are considered for further analysis. Also, the patients with cancer stage 0, 4 and Caucasian American (CA) patients with Hispanic ethnicity are excluded. Hence, the final population for analysis consisted of 3649 women out of which 2459 (67.39%) were CA and 1189 (32.61%) were AA.

The attribute "Survival Time Months" indicates the number of months the patient survived after diagnosis. The survival months of the patients are divided into four classes:

1. Class1 (0-2 years).
2. Class2 (2-5 years).
3. Class3 (5-9 years).
4. Class4 (> 9 years).

We used these four different categories as the class labels for evaluating the proposed differential predictive models. Most of the attributes were categorical with a finite set of values. Only 8 out of the 141 were numeric/continuous variables. (e.g. 'Survival Months', 'Age' etc). The coding used was the American Joint Committee on Cancer (AJCC) TNM edition 3 (1988) coding scheme to categorize patients by T stages (1, 2, 3, 4) and N stages (1, 2, 3). Cases were histologically grouped by labeling those with the International Classification of Diseases for Oncology (ICD-O-2) code. For building the continuum of trees, we have chosen a subset of most informative attributes to improve the performance of the models, to alleviate the effect of the curse of dimensionality [15], and to enhance generalization capability which also improves the interpretability of the model.

3.2. Disparity Analysis

We performed two different kinds of disparity analysis between the CA and AA population in the context of poverty and age. For measuring the disparity between the CA and AA population within different age groups, we divided the dataset into the following 4 age groups: (1) 20 – 50 years, (2) 50 – 65 years, (3) 65 – 75 years and (4) greater than 75 years. The intervals are chosen such that we can contrast women likely to pre-menopausal and post-menopausal stage, as well as the old women (>75 years).

For each patient in the dataset there were 141 attributes that describe different types of information about the patient. Based on the expert's suggestion we have selected the following 19 attributes which are grouped into four major categories:

1. *Demographic information:* "Age", "Marital Status", "Year".
2. *Tumor characteristics:* "T Value", "N Value", "M Value", "Tumor Size", "ER", "PR", "Histology", "Laterality", "Primary Site", "Regional Nodes positive", "Stage".
3. *Treatment:* "Radiation", "Surgery".
4. *Census Tract Socio-economic Status (SES):* "Median Household income", "Primary Payer", "Poverty".

Within each age group; while generating the trees for each of the racial groups, we varied the value of w to obtain a tree whose distance to the reference tree (in that age group) is the lowest. In our experiments, we have taken 100 different values for the parameter w ranging from 1.0 to 0.01 with decrements of 0.01. For any given group, accuracy at w=1 is equivalent to the accuracy of the tree induced by the standard C4.5 on that particular group. The accuracies for the models were calculated by 10 fold cross-validation. In Fig. 3, the deviation between maximum and minimum accuracy is not high (2.01% and 0.24% respectively). Hence, for each group, any of these 100 tree models (corresponding to 100 w values) can be regarded as a good representation of the data.

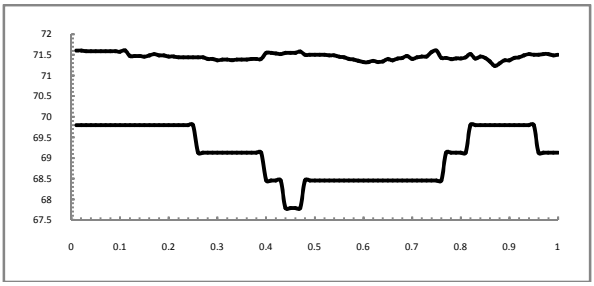

Fig 3: Parameter w vs. accuracy for both groups (dashed-CA, solid-AA) in Age group 1. w is on the x axis and the accuracy (in percentage) is on the y axis.

Fig. 4 plots different tree distances from the reference tree against different w values for the two groups (CA and AA) in age group 1 (20-50 years). During the tree generation process, the tree with the smallest distance from reference tree is considered to be the group model. We found the optimal CA model at w = 0.25. At this w, the distance of this tree with the reference tree (tree model for both CA and AA data in the age group 1) is 0.322. The optimal AA model is found at w = 0.41 and the corresponding distance is 0.46. We chose these two models and calculated the distance between them using our algorithm.

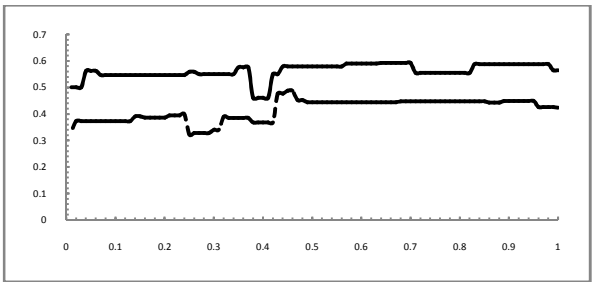

Fig 4: Parameter w vs. tree distance for both groups (dashed-CA, solid-AA) in age group 1. w is on the x axis and tree distance is in the y axis.

From *Table 1,* we see that the distance between the racial groups is highest in age group 3 (65-75 years) and is the lowest for age group 4 (>75 years). It means, in age group 3 the characteristic of CA and AA patient groups is most dissimilar while in age group 4 it is most similar. *Table 1* also lists the p-values. We used the *t* test to calculate the two sided p-values at the significance level of 0.05. For each group, the numerical values of the survival months are used to calculate the p-value. p-value is the probability of observing a difference as large as has been observed under the null hypothesis [16]. In our test, the null hypothesis is that the samples are drawn from the same population. From *Table 1,* we see that all the p-values except the one for age group 4 are statistically significant (e.g. smaller than the significance level of 0.05). In other words, the observed difference in age group 4 is not statistically significant. The observed difference as large as it is in age group 3 has a probability of 4.9272×10^{-4}, which is greater than that of all other age groups. This is another evidence for the two racial groups being most dissimilar in age group 3. However, in order to determine the difference in the data distribution, analysis based on p-value is not reliable. Because, p-value captures the differences in the data distribution with respect to only one attribute. In this case, it is the survival month. But, we are interested in a multivariate comparison. Moreover, many statisticians are not excited about such interpretation of p-values (e.g. the concept of 'significant' and 'very significant' results by comparing the magnitude of p-values). They consider every result as either statistically significant or not statistically significant at a particular significance level. *The novelty of our approach is that it captures the disparity of groups through comparison of multivariate models*.

Table 1: Distances between the CA and AA models in different disparity groups.

Disparity	Groups	Distance	p-value
Age	20-50 years	0.642328	3.3042×10^{-6}
	50-65 years	0.504311	4.1244×10^{-9}
	65-75 years	0.662227	4.9272×10^{-4}
	>75 years	0.446609	>0.05
Poverty	<0.07	0.569196	0.0376
	>0.07	0.3881	8.7781×10^{-5}

While comparing the two trees corresponding to the two racial groups in age group 4, it is found that there are 6 attributes that appears either in the tree model of CA or in the tree model of AA, and not in both. These are: "Poverty", "Marital Status", "T", "Surgery", "Median Household Income" and "N". These are the most responsible attributes for the disparity. To verify if this is a significant result, a decision tree model is learnt for age group 4 taking the "Race" attribute as the class label. This tree has 4 attribute levels. "Median Household Income" appears in the root and both of its children are the attribute "Poverty" appearing in the next level. Thus, *besides finding out the extent of different behavior of the groups in different regions; our algorithm is also capable of pointing out the attributes responsible for such differences*.

A similar analysis in the context of poverty gives us the results listed in *Table 1*. The two racial groups are more dissimilar in poverty < 0.07 than in poverty > 0.07. 0.07 is the median of the poverty values in the data. Hence, poor CA and AA patients are more similar in terms of their survival model. They are more dissimilar when they are not poor. It indicates that the lower survivability of AA population can not be attributed only to the poor economic status. Even in affluent economic conditions they behave significantly different.

4. Conclusion

In this paper, we propose a new method of differential predictive modeling and propose a solution to the problem by developing a continuum of tree models. We demonstrate the usefulness of this new approach by modeling the racial survival disparities in the breast cancer data. Since the problem posed in this paper has not been studied in the literature, there is a great potential to build other reliable solutions. We can use other metrics like entropy or misclassification error rate as the split criteria and make a comparative study of the suitable split measures. We also plan to study other classification and regression models in this framework.

Acknowledgments

The authors would like to thank Elayne Arterbery and Fawn Vigneau of Karmanos Cancer Institute, Detroit for useful discussions.

References

[1] J. R. Quinlan, “C4.5: Programs for Machine Learning”, Morgan Kaufmann Publishers Inc., 1993.

[2] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, “Classification and Regression Trees”, Wadsworth, Belmont, CA, 1984.

[3] T. Hastie, R. Tibshirani and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer-Verlag, Second Edition, 2009.

[4] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. Yu, Z. Zhou, M. Steinbach, D. J. Hand and D. Steinberg, “Top 10 algorithms in data mining”, Knowledge and Information Systems. Vol 14, No 1, pp 1-37, 2008.

[5] Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques”, Second Edition, Morgan Kaufmann, 2006.

[6] M Wood, C Johnson et al. “Race and Differences in Breast Cancer survival in a Managed Care Population”, Journal of the National Cancer Institute, Vol. 91, No 17, pp 1487- 1491, 1999.

[7] M.S. Simon, M. Banerjee, H. Crossley, F.D. Vigneau, A. Noone and K.L. Schwartz, “Racial differences in breast cancer survival in the Detroit Metropolitan area”, Breast Cancer Research and Treatment, Vol. 97, pp 149–155, 2006.

[8] National Cancer Institute (NCI): Cancer Control and Population Sciences, National Institutes of Health, “Overview of Health Disparities Research: Health Disparities Definition”, Accessed on July 20, 2009 at: http://cancercontrol.cancer. gov/od/hd-overview.html.

[9] J. Goldberg, W. Hayes and J. Huntley, "Understanding Health Disparities", Health Policy Institute of Ohio, pages 4-5, November 2004.

[10] American Public Health Association (APHA), Eliminating Health Disparities: Toolkit, 2004.

[11] Kamangar, Farin, Dores, M. Graca, Anderson and F William, “Patterns of Cancer Incidence, Mortality, and Prevalence Across Five Continents: Defining Priorities to Reduce Cancer Disparities in Different Geographic Regions of the World”, Journal of Clinical Oncology, Vol. 24, pp 2137-2150, 2006.

[12] J. Bigby and M. Holmes, “Disparities across the breast cancer continuum”, Cancer Causes and Control 16:35-44, 2005.

[13] Kathy S. Albain, Joseph M. Unger, John J. Crowley, Charles A. Coltman Jr and Dawn L. Hershman, “Racial Disparities in Cancer Survival Among Randomized Clinical Trials Patients of the Southwest Oncology Group”, Journal of the National Cancer Institute Advance Access, Vol. 101, pp 984-992, July 15, 2009.

[14] American Cancer Society, “Cancer Facts and Figures 2008”. Accessed at www.cancer.org on July 20, 2009.

[15] Richard O. Duda, Peter E. Hart and David G. Stork, “Pattern classification”, 2nd edition, Wiley, New York, 2001.

[16] John A. Rice, “Mathematical Statistics and Data Analysis”, 3rd edition, Duxbury Press, 2007.