Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 12, December 2014, pg.584 – 589

RESEARCH ARTICLE

WHO ARE the LIKELIEST SURVIVORS of BREAST CANCER- An ANALYSIS by K NEAREST NEIGHBORS ALGORITHM

Md. Ahasan Uddin Harun

[1]*Department of Mathematics & Statistics, Sam Houston State University, USA*

harunahasan@gmail.com

Abstract— As deaths caused by breast cancer have become a major concern, researchers are trying to design & implement algorithms so that preventive measures can be adopted in an effective way. Hence, the aim of this study is to improve prediction of survivability of breast cancer with the application of K-nearest neighbors algorithm and its corresponding ensemble version. WEKA software has been used for this analysis. The findings of our study confirm that such approach provides effective prediction from statistical standpoint and thus use of such classification procedure may be adopted, contributing to improved diagnosis in breast cancer.

Keywords— Data mining, Classification, K-Nearest Neighbors, WEKA, Breast cancer

I. INTRODUCTION

Since the advancement in medical field depends on the effective analysis of large data, statistical learning techniques are commonly used to discover consistent and meaningful pattern. These techniques are based on inductive inference where a phenomenon is observed and then a model is built for that phenomenon so that predictions can be made using this model.

It is possible to predict the outcome of a certain disease by extracting information from the database related to that disease. In this paper, we are going to use a classification algorithm known as K-Nearest Neighbors (KNN) and its ensemble counterparts to predict survivability of breast cancer as data driven such approach is becoming a trend in many scientific areas such as medicine.

This paper has been designed as follows: in the following 2 sections insight into related works and description of data set have been presented. Then research methodology which is going to be followed in this study will be discussed. Next, an attempt will be made to summarize & highlight the obtained results. In the final phase, limitation of the study and scope of further improvement will be illustrated.

II. RELATED WORKS

We are going to start our study with a thorough look of how other researchers approached such dataset. This will help us to have a better comprehension of the data and we will be able to build foundation of our study as the other researchers might have proposed effective methods that are worth investigating.

Delen et al.[5] obtained a survival prediction with 93% accuracy using a breast cancer database consisting of 202,932 records. Liu Ya-Qin[8] used j48 algorithm and its bagged ensemble version for such prediction. Chaurasia et al.[6] investigated 3 data mining techniques RepTree, RBF network and simple logistic approach and they found that simple logistic algorithm performs the best with 74.47% accuracy.

III. DATA UNDERSTANDING

The source of the breast cancer data is University Medical Center, Institute of Oncology, Ljubljana, Slovenia .There are 286 observations & 10 attributes of which 1 is class attribute. If a closer look is taken at the data set, we see that the class attribute is nominal and has binary classification. Rest of the attributes are also nominal. There are very few missing values and number of distinct values is also few. This has been illustrated in table 1 and histograms of all the attributes have been provided in fig 1 from which we can see that no apparent structure exists in the data set. Attributes such as tumour size, deg-malig have normal distribution whereas inv-nodes, breast-quads & menopause seem to have exponential distribution.

To get a better insight into this dataset, we will transform it to get a new one where missing values will be removed. Another version of dataset will also be created where missing values will be replaced with majority values of the attributes. These 3 dataset versions-unaltered & altered datasets will be the basis of our investigation in this study.

TABLE I
MISSING & DISTINCT VALUES OF ATTRIBUTESS

Attribute name	Missing values	Distinct values
age	0	6
menopause	0	3
tumor-size	0	11
Inv-nodes	0	7
node-caps	8	2
def-malig	0	3
breast	0	2
breast-quad	1	5
irradiat	0	2

IV. RESEARCH METHODOLOGY

In this section, the research methodology followed in this study is going to be presented.

A. Experimental setup

In this study, WEKA toolkit (version 3.6.11) has been used for analysis. It is the product of the University of Waikato (New Zealand) and in its modern form was first implemented in 1997.It has been written in JAVA language and uses GNU general public license (GPL).WEKA contains a GUI for interacting with data files and creating exploratory results. Moreover, it also provides access to SQL database and is able to process the result retrieved by a database query.

B. Overview of K-Nearest Neighbors Algorithm (KNN))

We are going to investigate performances of K-Nearest Neighbors Algorithm (KNN) on the 3 versions of dataset. In WEKA, this algorithm is known as IBK (instance based learner) .

Instead of creating a model, IBK generates a prediction for testing observation just in time. The IBK algorithm uses a distance measure to locate k 'close' observations in the training data for each test observations and uses those selected observations for prediction.

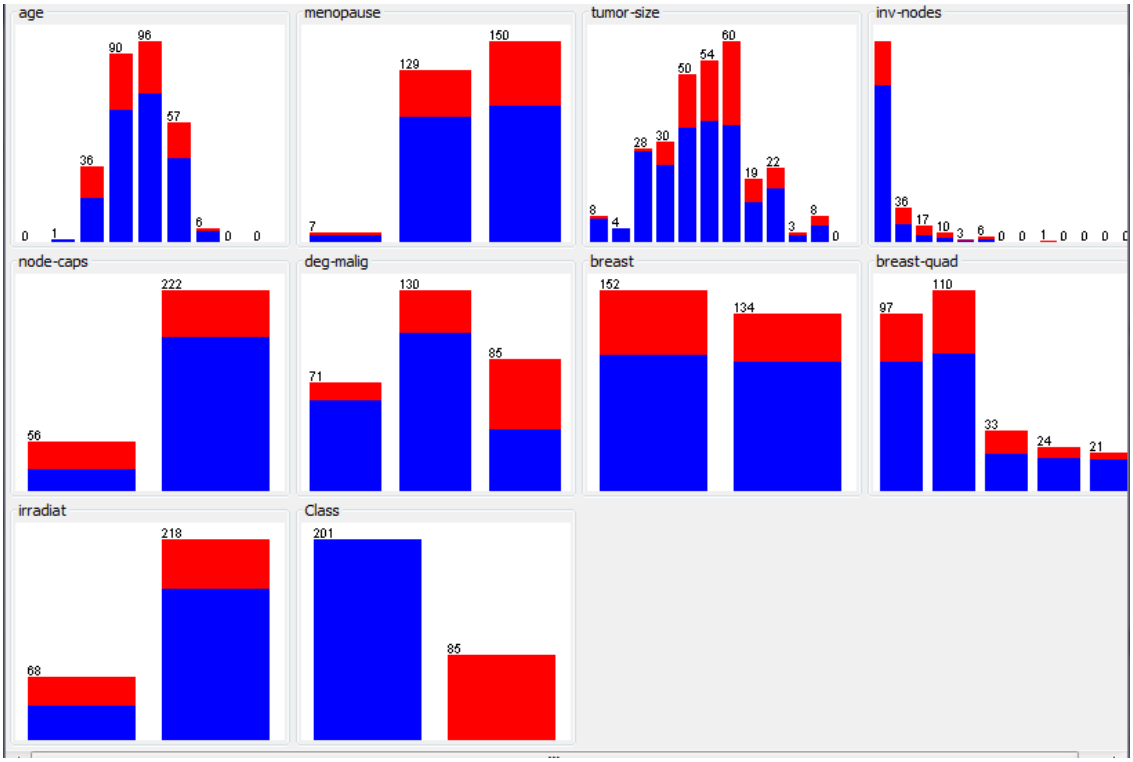

Fig. 1 Histogram of the attributes in Breast Cancer data set

In this experiment, we are curious to locate which of 3 distance measures –Manhattan, Chebyshev & default Euclidean distance gives better prediction. Therefore, we will tune the parameters of IBK in WEKA accordingly to facilitate our experiment.

C. Ensemble approach

In this approach, results of several methods are combined to achieve better performance. If several 'good enough' models specialize in various segments of the problem under study, then ensemble procedure performs quite satisfactorily.

Several ensemble procedures are available in data mining literature for analysis purpose. Of these, boosting method [7] will be applied in this study. In boosting procedure, at first a base classifier is prepared on training data. Then a second classifier is used to create new models that target the observation in the training data that first classifier got wrong. The process is continued to add classifiers so long as a threshold is obtained. In WEKA, boosting has been provided by Adaboost M1 algorithm.

Since ensemble procedure is far more complex than traditional methods and traditional methods give a good base level from which it is possible to improve and create new ensembles, it is customary to use ensemble procedure after more traditional methods are exhausted. Following this practice, after exploring the dataset through K-Nearest Neighbors algorithm, boosted versions of it will be investigated.

A. Cross-validation

For this study, principle of 10 fold cross-validation will be used where at first, 10 equal sized data sets will be created from given data. Then each data set will be partitioned into 2 groups- 90% for training & 10 % for testing. After that, a classifier will be produced with an algorithm from 90% labelled data and applied to the 10% testing data for set 1. This procedure will be continued for set 2 through 10. In the final phase, performance of the classifiers created from 10 equal sized (training & testing) sets will be averaged.

B. hypothesis & test statistic

The goal of this study is to compare the performances of K-Nearest Neighbors algorithm (KNN) and its corresponding ensemble version across unaltered & pre-processed data sets for predicting survivability of breast cancer patients. For this study, the null hypothesis that KNN with all 3 distance measures & their ensemble versions do equally well has been set up .

The statistic which will be used in testing this hypothesis is 2 sided corrected paired t test as standard t- test can produce too many significant differences because of dependencies in the estimates [10] . We will use this statistic at 95% confidence level.

V. RESULT ANALYSIS

At first, we are curious to see which algorithm is the best. It is possible to do so by ranking the algorithms by the number of times a particular algorithm beats others. In WEKA Experimenter, this can be done by clicking 'select' button for the 'test base' and then choosing 'ranking' and clicking 'perform test' button. Rankings

TABLE 2
RESULTS OF ALGORITHM RANKING

METHODS	CRITERIA								
	ACCURACY			F MEASURE			ROC		
	wins	loses	overall	wins	loses	overall	wins	loses	overall
IBK (Euclidean)	9	0	9	8	0	8	6	0	6
IBK (Manhattan)	9	0	9	8	0	8	6	0	6
IBK (Chebyshev)	2	0	2	2	0	2	0	12	-12
Boosted IBK (Euclidean)	2	6	-4	2	6	-4	6	0	6
Boosted IBK (Manhattan)	2	6	-4	2	6	-4	6	0	6
Boosted IBK (Chebyshev)	0	12	-12	0	10	-10	0	12	-12

TABLE 3
RESULTS FROM ACCURACY PERSPECTIVE

DATASETS	METHODS					
	IBK (Euclidean)	IBK (Manhattan)	IBK (Chebyshev)	Boosted IBK (Euclidean)	Boosted IBK (Manhattan)	Boosted IBK (Chebyshev)
Unaltered dataset	72.85 (6.93)	72.85 (6.93)	70.30 (3.67)	69.76 (7.90)*	69.76 (7.90)*	60.14 (16.57)*
Dataset (removed missing values)	74.02 (6.62)	74.02 (6.62)	70.85 (3.41)	71.02 (7.45)*	71.02 (7.45)*	33.72 (4.69)*
Dataset (replaced missing values)	72.74 (7.00)	72.74 (7.00)	70.30 (3.67)	69.66 (7.95)*	69.66 (7.95)*	35.37 (6.12)*

from table 2 illustrate the number of statistically significant wins each algorithm has against all other competing algorithms. A win means a performance that is better than that of other algorithm and that the difference is statistically significant

From accuracy standpoint, IBK with Manhattan & Euclidean distance perform best across all data sets. Interestingly, boosted versions did not perform satisfactorily. From F measure standpoint; these 2 again performed the best. Boosted IBK with Chebyshev distance was the worst than the other tuned or non tuned versions. And finally, from ROC standpoint, interestingly, both IBK with Chebyshev distance and its boosted counterpart performed the worst..

Now we want to see what scores these algorithms actually will achieve. To do so, each version of IBK algorithm will be run 10 times on the data set. The reported result is the mean and the number in brackets is the standard deviation of those 10 runs. Significant results have been indicated by * symbol.

From accuracy perspective, from table 2, we observe that dataset where observations with missing values were removed gives better prediction. IBK with Manhattan & Euclidean distance provided the highest mean accuracy at 74.02% with 6.62% standard deviation. Both of these significantly outperformed not only corresponding boosted versions but also boosted version of IBK with Euclidean distance, though superiority over IBK with Chebyshev distance was not supported from statistical significance framework

Similarly, from F metric viewpoint, almost similar results were obtained which have been shown in table 4. However, from ROC point of view, presented in in table 5, we see that though superiority of IBK with Manhattan & Euclidean distance in case of data set with removed missing values were proved from statistical significance point of view over boosted version IBK with Chebyshev distance that was not the case over boosted IBK with Euclidean & Manhattan distance. But their superiority was established over IBK with Chebyshev distance and that was statistically significant.

So it is apparent that IBK with Manhattan distance variation performs at the top level. That is encouraging as it looks like a configuration that performs equally well to the algorithm default (i.e. Euclidean distance) has been found. Moreover, boosted versions performed very poorly in all 3 metrics. It seems that limits of IBK algorithm have been pushed to the limit.

TABLE 4
RESULTS FROM F MEASURE PERSPECTIVE

DATASETS	METHODS					
	IBK (Euclidean)	**IBK (Manhattan)**	**IBK (Chebyshev)**	**Boosted IBK (Euclidean)**	**Boosted IBK (Manhattan)**	**Boosted IBK (Chebyshev)**
Unaltered dataset	0.82 (0.05)	0.82 (0.05)	0.82 (0.02)	0.79 (0.06)*	0.79 (0.06)*	0.63 (0.30)
Dataset (removed missing values)	0.83 (0.04)	0.83 (0.04)	0.82 (0.02)	0.80 (0.05)*	0.80 (0.05)*	0.15 (0.10)*
Dataset (replaced missing values)	0.82 (0.05)	0.82 (0.05)	0.82 (0.02)	0.79 (0.06)*	0.79 (0.06)*	0.18 (0.13)*

TABLE 5
RESULTS FROM ROC PERSPECTIVE

DATASETS	METHODS					
	IBK (Euclidean)	**IBK (Manhattan)**	**IBK (Chebyshev)**	**Boosted IBK (Euclidean)**	**Boosted IBK (Manhattan)**	**Boosted IBK (Chebyshev)**
Unaltered dataset	0.64 (0.10)	0.64 (0.10)	0.54 (0.06)*	0.61 (0.09)	0.61 (0.09)	0.54 (0.06)*
Dataset (removed missing values)	0.65 (0.11)	0.65 (0.11)	0.54 (0.06)*	0.63 (0.10)	0.63 (0.10)	0.53 (0.06)*
Dataset (replaced missing values)	0.64 (0.10)	0.64 (0.10)	0.54 (0.06)*	0.61 (0.09)	0.61 (0.09)	0.54 (0.06)*

VI. LIMITATION OF STUDY & SCOPE OF IMPROVEMENT

There are some limitations of this study. Results which have been obtained here may be limited to the country or the institution from which observations were collected. Moreover, obtained results may also be limited for the time frame data set was collected. Furthermore, dataset used in this study is quite small, which may hinder performance of the algorithm under study.

Nonetheless, as a starting point this dataset can be used to gain better understanding of the survivability of breast cancer patients. And surely, there are scopes of further investigation. For example, in this analysis only KNN and its boosted version have been used. In the future work, other data mining techniques can be introduced to gain a better understanding of the dataset. Moreover, to find a better prediction model, other ensemble approaches such as bagging and stacking can be tested to compare the prediction results. Furthermore, 'feature engineering' i.e. attribute decomposition & aggregation can be done to investigate whether useful information can be extracted from the dataset under study.

VII. CONCLUSIONS

The goal of this study was to use K-Nearest Neighbors algorithm & its ensemble counterpart for better prediction of survivability of breast cancer patients. To do so, preprocessing of data was done. Then both altered & unaltered data sets were used so that we can understand how different data sets affect the results. The results indicate that K-Nearest neighbors with Manhattan & Euclidean distance generally give better performance.

In doing so, we have analyzed their performance by 3 different metrics. There is no doubt that a large dataset would provide more robustness in analysis. Nonetheless, this study will give a better understanding of KNN classification technique in medical data analysis as the results have been justified from statistical framework.

REFERENCES

[1] M. Kuhn and K. Johnson, Applied Predictive Modelling, 1st ed., Berlin, Germany: Springer-Verlag , 2013
[2] T. Hastie, R Tibshirani and J. Friedman, The Elements of Statistical Learning, 2nd ed., Berlin, Germany: Springer-Verlag, 2009
[3] N. Zumel and J. Mount, Practical Data Science with R, 1st ed., Connecticut , USA: Manning Publications Co.,2014
[4] I. H. Witten , E. Frank and M. A. Hall, Data Mining :Practical machine learning tools & techniques, 3rd ed. , Massachusetts ,USA: Morgan Kaufmann ,2011
[5] D. Delen, G. Walker,A. Kadam, "Predicting Breast cancer survivability : a comparison of 3 data mining methods" , Artificial Intelligence in Medicine, vol. 34, pp 113-127, Jun. 2005.
[6] V. Chaurasia , S. Pal, " Data mining techniques: to predict and resolve breast cancer survivability" , International Journal of Computer Science and Mobile Computing, vol. 3 pp 10-22, Jan. 2014
[7] R. E. Schapire, (1990). "The Strength of Weak Learnability" Boston, USA: Kluwer Academic Publishers, 1990
[8] Liu Ya-Qin,W. Cheng, Z. Lu , "Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data" ,3rd International Conference on Bioinformatics and Biomedical Engineering, 2009
[9] S. Haykin , Neural Networks : a comprehensive foundation, 1st edition, London, Prentice Hall, 1999
[10] T. Dietterich , "approximate statistical tests for comparing supervised classification learning algorithms" , Neural Computation, vol. 10, pp1895-1924, 1998
[11] V. K. Mago and N. Bhatia, Cross-disciplinary Applications of Artificial Intelligence and Pattern Recognition : Advancing Technologies, 1st ed., Pennsylvania, USA : IGI Global, 2011

