ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 1, pp: (96-99), Month: April 2015 – September 2015, Available at: www.paperpublications.org

A Study on Cancer Perpetuation Using the Classification Algorithms

ANITA KUMAR

Bishop Heber college Trichy-17, India

Abstract: **Analysis of cancer datasets is one of the important research in data mining techniques. In the present work, classification techniques such as CART, Random Forest, LMT, and Naive Bayesian are used. The result predicts that Random forest method using training dataset outperforms the remaining methods. The random forest method using training dataset have less value of absolute relative error. Relative absolute error of LMT is high for cancer survival dataset. Value of absolute relative error is greater than 50% for almost all the algorithms except for random forest method using training dataset.**

Keywords: **Classification techniques- CART, Random Forest, LMT, and Naive Bayesian.**

1. INTRODUCTION

Data mining is the process of analysing data from different perspectives and summarizing it into important information so as to identify hidden patterns from a large data set. Researchers in many fields have shown great interest in data mining.

Cancer also known as a malignant tumor or malignant neoplasm, is a group of diseases involving aberrant cell growth with the potential to invade or spread to other parts of the body.[1][2] it does not mean that all tumors are cancerous; benign tumors do not spread to other parts of the body. The Possible signs and the symptoms include: a new lump, abnormal bleeding, cough for a very long period, unexplained weight loss, among others.[3] While these symptoms might indicate cancer, they may also occur due to other complications. There are over 100 different types of known cancers that affect humans.

In bioinformatics age, cancer datasets can be used for the cancer diagnosis and treatment, which can improve human aging [4].The Data mining techniques, such as the pattern association, classification and clustering, are mostly applied in the cancer and gene expressions correlation studies. Bioinformatics provides logic for developing novel data mining methods.

Classification of datasets based on a predefined knowledge of the objects is a data mining [5].Knowledge management technique is used in grouping the same data objects together. The ultimate goal of a supervised learning algorithm is to build a classifier that can be used to classify unlabelled instances accurately [6]. Data classification contains supervised learning algorithms as it assigns class labels to data objects based on the relationship between the data sets with a pre-defined class label. Classification algorithms have a very wide range ofapplications like fraud detection, churn prediction, artificial intelligence, neural networks and the credit card rating etc. [7]. There are many classification algorithms available in literature and is a well studied area in data mining. Numerous classification algorithms have been proposed in the literature, such as classification and regression tree [8], Logistic Model Tree [9], [10], Random forest [11], Bayesian classifiers [12]. Cancer detection is one of the most important research topics in biomedical science. Biomedical research applies a wide range of designs to solve problems in laboratory, clinical, and population settings [13]. Here in this paper we studied various classification algorithms like CART, Random Forest, LMT and Naïve Bayesian over different cancer survival dataset. Accuracy is the main objective to estimate the performance of these algorithms over cancer datasets.

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 1, pp: (96-99), Month: April 2015 – September 2015, Available at: www.paperpublications.org

2. METHODOLOGY

A study of Cancer Surveillance using Data Mining, and the Decision Support Systems can reduce the national cancer burden or the oral complications of cancer therapies. Here, in this paper, we study various classifications of algorithms like CART, Random Forest, LMT and Naïve Bayesian over different cancer survival dataset. The data explored in this research was obtained from the Dataset available in the UCI . Patients with highly developed cancers of the stomach, bronchus or colon were treated with acrobat. The rationale of this study is to resolve if the survival times differ with respect to the organ affected by cancer. There were no missing values and the dataset was complete. The main aim of processing the data is to discriminate cancer survivability in people with a two-decision classification problem.

2.1 CART:

A classification and regression tree (CART) is a recursive and gradual refinement data mining algorithm of building a decision tree. CART algorithm is widely used statistical procedure based on tree structure that can produce classification and regression trees, depending on the dependent variable weather it is categorical or numeric, respectively and generates binary tree.

2.2 LMT:

A Logistic Model Tree (LMT) is an algorithm for supervised learning tasks which is combined with linear Logistic regression and tree induction. LMT creates a model tree with a standard decision tree structure with logistic regression functions at leaf nodes. In LMT, leaves have an associated logic regression function instead of just class labels.

2.3 Random forest:

Random forest is an ensemble classifier that consists of many decision tree, and outputs the class that is the mode of the class's output by individual trees. The Random Forests grows many classification trees without pruning. Then a test sample is classified by each decision tree and random forest assigns a class which have maximum occurrence among these classifications.

2.4 Naïve Bayesian:

Naïve Bayesian classifier is a simple probabilistic classifier based upon Bayes theorem with strong (naive) independence assumptions. Naïve Bayesian classifier is based on Bayes conditional probability rule and is used for performing classification tasks. All attributes of the dataset are considered independent of each other. In general, a naïve Bayes classifier assume that the presence (or absence) of a selective feature of a class is unrelated to the presence (or absence) of any other feature. An advantage of the naïve Bayes classifier is that it rebuild amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

TABLE 1

S.NO	ALGORITHM	CORRECTLY CLASSIFIED	INCORRECTLY CLASIFIED	ABSOLUTE RELATIVE ERROR
1.	CART	36%	64.06%	91.25%
2.	LMT	34.37%	65.7%	96.06%
3.	RANDOM FOREST	42.18%	57.9%	79%
4.	NAIVE BAYESIAN	39.06%	61%	79%

TABLE 2

S.NO	ALGORITHM	CORRECTLY CLASSIFIED	INCORRECTLY CLASIFIED	ABSOLUTE RELATIVE ERROR
1.	CART	58%	42%	74.67%
2.	LMT	45%	55%	91.4%
3.	RANDOM FOREST	93.75%	6.25%	21.4%
4.	NAÏVE BAYESIAN	41%	59%	88.1%

3. RESULTS AND DISCUSSION

Study of cancer survival dataset is also done in (Table 1). Here Random forest algorithm outperforms all other classification algorithms used in the study. Comparison of the classification techniques which includes CART, Random Forest, LMT, and the Naive Bayesian over different cancer survival dataset shows that Random forest method using training dataset outperforms the other methods (Table 2). Relative absolute error of LMT is high for cancer survival dataset. Value of absolute relative error is greater than 50% for almost all the algorithms. Only the random forest method using training dataset have less value of the absolute relative error.

4. CONCLUSION

On Comparing the o classification techniques on cancer survival dataset including the Random Forest ,CART, LMT and Naïve Bayesian ,it is clear that Random forest method outperforms the remaining methods. Absolute relative error for the algorithm (Random Forest) is also less than the Absolute relative error of the other algorithms.

REFERENCES

[1] "Cancer Fact sheet ". World Health Organization. February 2014. Retrieved on 10 June 2014.

[2] "DefiningCancer". National Cancer Institute. Retrieved on 10 June 2014.

[3] "Cancer - Signs and symptoms". NHS Choices. Retrieved on 10 June 2014.

[4] Christoph Bock, Thomas Lengauer, “Computational epigenetic,” Bioinformatics, Vol. 24, No.1, pp. 1-10, in the year 2008.

[5] Yi Peng, Gang Kou, Yong Shi, Zhengxin Chen, “A descriptive framework for the field of data mining and Knowledge discovery,” Vol. 7, No. 4, pp. 639-682, in the year 2008.

[6] H. Friedman, R. Kohavi, Y. Yun, “Lazy decision trees,” In Proceedings of the Thirteenth National Conference on Artificial Intelligence, AAAI Press and the MIT Press, pp. 717-724, in the year 2006.

[7] Richard J. Bolton, David J. Hand, “Statistical Fraud Detection: A Review,“ Statist. Sci., Vol. 17, No. 3, pp. 235-255, in the year 2002.

[8] Breiman L, Friedman J, Olshen R, Stone C, "Classification and Regression Trees," Wadsworth International Group, in the year 2004.

[9] Frank E, Wang Y, Inglis S, Holmes G, Witten I. H, “Using model trees for classification,” Machine Learning, Vol. 32, No. 1, pp. 63–76, in the year 2008.

[10] Niels Landwehr, Mark Hall, Eibe Frank, "Logistic Model Trees," Machine Learning, Vol. 59, No. 1-2, pp.161-205, in the year 2005.

[11] Leo Breiman, "Random Forests," Machine Learning, Vol. 45, No. 1, pp. 5-32, 2001.

[12] Langley P, Iba W, Thompson K, "An analysis of Bayesian classifiers," In Proceedings of AAAI-92, AAAI Press, pp. 223-228, in the year 2002.

[13] John C. Bailar, Thomas A. Louis, Philip W. Lavori, Marcia Polansky, "A Classification for Biomedical Research Reports," N Engl J Med, Vol. 311, No. 23 pp. 1482-1487, in the year 2010.

[14] Golub T.R, Slonim D.K, Tamayo P, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science. Vol. 286, No.5439, pp. 531–537, in the year 2009. 343International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 2, No. 2, April 2011.

[15] Alizadeh A, Eisen M.B, Davis R.E, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC,Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," Nature, Vol. 403, No. 6769, pp. 503–511, in the year 2000.

[16] Nielsen T.O, West R.B, Linn S.C, Alter O, Knowling MA, O'Connell JX, Zhu S, Fero M, Sherlock G, Pollack JR, Brown PO, Botstein D, van de Rijn M,"Molecular characterisation of soft tissue tumors: a gene expression study," Lancet, Vol. 359, No. 9314, pp. 1301-1307, in the year 2012.

[17] Thangaraju, P., and G. Barkavi. "Lung Cancer Early Diagnosis Using Some Data Mining Classification Techniques: A Survey." COMPUSOFT, An international journal of advanced computer technology (IJACT) 3.6 (2014).

[18] Krishnaiah, V., Dr G. Narsimha, and Dr N. Subhash Chandra. "Diagnosis of lung cancer prediction system using data mining classification techniques."International Journal of Computer Science and Information Technologies 4.1 (2013): 39-45.

[19] Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5.5 (2013): 241-266.

[20] Ramachandran, P., et al. "Cancer Spread Pattern–an Analysis using Classification and Prediction Techniques." Cancer 2.6 (2013).

[21] Yang, Chun-Yi. "A Hybrid of Data Mining and Statistical Analysis Approach on Association between Pulmonary Tuberculosis and Lung Cancer." (2014).

[22] Ada, Rajneet Kaur. "Using Some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient." (2013).

[23] Khedr, Aymn E., and Abd El-Ghany AM Mohmed. "A proposed image processing framework to support Early liver Cancer Diagnosis." Life Sci J 9.4 (2012): 3808-3813.

[24] Halder, Subhas. An Approach to Diagnosis of Cancer using k-nearest Neighbor (k-NN) Algorithm. Diss. JADAVPUR UNIVERSITY KOLKATA, 2013.

[25] Ramachandran, P., N. Girija, and T. Bhuvaneswari. "Early Detection and Prevention of Cancer using Data Mining Techniques." International Journal of Computer Applications 97.13 (2014): 48-53.