Predicting Cancer Survivability Using Classification Algorithms

DSVGK Kaladhar[1], B. Chandana[2] and P. Bharath Kumar[3]

[1,2,3]Department of Bioinformatics, GIS,
GITAM University, Visakhapatnam-530045, India

***Abstract*:** Analysis of cancer datasets is one of the important research in data mining techniques. In the present work, data is obtained from the Oxford Dataset and classification techniques such as CART, Random Forest, LMT, and Naïve Bayesian were used. The result predicted 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 algorithms, Cancer survival dataset**.**

1. Introduction

Data mining (or data discovery) is the process of analyzing and predicting data from many different dimensions and summarize the relationships identified [1]. Data mining softwares has a number of analytical tools for analyzing information from different perspectives like database systems and machine learning, summarizing it into useful information that can be used to increase accuracy of the data. Researchers in many fields have shown great interest in data mining [2].

Data mining in Cancer research is one of the important research topics in biomedical science [3]. In bioinformatics age, cancer datasets can be used for the cancer diagnosis and treatment, which can improve human aging [4]. Data mining techniques, such as pattern association, classification and clustering, are now frequently applied in 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 similar data objects together. The main 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 wide range of applications like fraud detection, churn prediction, artificial intelligence, neural networks and 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].

Classification of cancer types using gene expression datasets has been considered by Golub [14], Alizadeh [15] and Nielsen [16]. Fort et al., 2005 is proposed a new classification method, combining partial least squares (PLS) and Ridge penalized logistic regression [17].

Classification is very important among the techniques of data mining [18]. 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.

2. Methodology

A study of Cancer Surveillance using Data Warehousing, Data Mining, and Decision Support Systems [19] can reduce the national cancer burden or the oral complications of cancer therapies. Here in this paper, we studied various classification algorithms like CART, Random Forest, LMT and Naïve Bayesian over different cancer survival dataset. In first few sections we briefly described these algorithms and after that an overview of cancer datasets is given. Results are discussed in the last section of the paper.

The data explored in this research was obtained from the Oxford Cancer Survival Dataset [20]. Patients with highly developed cancers of the stomach, bronchus, colon, ovary or breast were treated with ascorbate. The rationale of the study was to resolve if the survival times differ with respect to the organ affected by the 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.

A classification and regression tree (CART**)** is arecursive and gradual refinement data mining algorithm of building a decision tree. CART algorithm [8] is widely used statistical procedure based on tree structure that can produce classification and regression trees, depending on whether the dependent variable is categorical or numeric, respectively and generates binary tree.

A Logistic Model Tree (LMT) [10] 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 a associated logic regression functions instead of just class labels.

Random forest [11] 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 algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler. 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.

Naïve Bayesian classifier [12] 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 naive 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 naive Bayes classifier is that it rebuild amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

3. Results and Discussion

Study of cancer survival dataset is also done using 10-fold cross-validation (Table 1). Here Random forest algorithm outperforms all other classification algorithms used in the study.

Comparison of the classification techniques including CART, Random Forest, LMT, and Naïve Bayesian over different cancer survival dataset shows that Random forest method using training dataset outperforms the remaining 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 absolute relative error.

Tree generated by CART and naïve Bayesian Classifier for the classification of dataset is shown in Figure 1 and Figure 2. Bronchus and colon datasets has high weight sum and ovary has least weight sum shown in Naïve Bayesian classifier.

Table 1. Classification for Cancer Survival Dataset using 10 fold cross validation

S.no	Algorithm	Data at 10 fold cross validation		
		Correctly Classified	Incorrectly Classified	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. Classification for Cancer Survival Dataset using training dataset

S.no	Algorithm	Data using Training dataset		
		Correctly Classified	Incorrectly Classified	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.	NAIVE BAYESIAN	41%	59%	88.1%

```
=== Classifier model (full training set) ===

CART Decision Tree

Survival < 628.0
|  Survival < 247.0
|  |  Survival < 57.0
|  |  |  Survival < 41.0
|  |  |  |  Survival < 22.0: Bronchus(1.0/1.0)
|  |  |  |  Survival >= 22.0: Breast(2.0/2.0)
|  |  |  Survival >= 41.0: Stomach(4.0/0.0)
|  |  Survival >= 57.0: Bronchus(12.0/9.0)
|  Survival >= 247.0: Colon(9.0/7.0)
Survival >= 628.0: Breast(9.0/8.0)
```

Figure 1. Classification data model using CART Decision Tree

```
=== Classifier model (full training set) ===

Naive Bayes Classifier

                Class
Attribute       Breast  Bronchus     Colon     Ovary   Stomach
                (0.17)    (0.26)    (0.26)    (0.1)     (0.2)
============================================================
Survival
  mean        1400.0358  208.2122  460.2584  879.7268  286.6078
  std. dev.   1176.8533  205.3406  414.6904 1012.4234  330.8772
  weight sum         11        17        17         6        13
  precision     62.0984   62.0984   62.0984   62.0984   62.0984
```

Figure 2 Classification data model using Naïve Bayesian Classifier

Vital information has been revealed from the data of National Cancer Institute Surveillance, Epidemiology, and End Results program for 1985–1989. The data has shown that older persons have a risk of developing cancer 10 times greater than that for individuals younger than 65 [21]. According to Houston et al. [22] and Cios et al. [23], the data in statistical research is flattering a common complement to many scientific areas like medicine and biotechnology.

The Weka machine learning workbench provides environment for data mining problems in bioinformatics research. The software can provide automatic classification, regression, clustering and feature selection techniques on the same problem [24]. Medical data with mathematical understanding of estimation and hypothesis formation may be fundamentally different than those from other data collection methods.

The data will judge the effectiveness and improvement rate of an algorithm and can be successfully implemented in a complex problem domain. The submitted data for classification contains two attributes- survival and Organ. Based on classification techniques and confusion matrix analysis on cancer survival dataset, random forest method had predicted better results in comparison with CART, LMT and Naïve Bayesian methods.

4. Conclusion

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

Acknowledgement

Authors would like to thank management and staff of GITAM University, Visakhapatnam, India for their kind support in bringing out the above literature and experimentation.

References

[1] Donna K. Slonim, “From patterns to pathways: gene expression data analysis comes of age,” Nature Genetics supplement, Vol. 32, pp. 502 – 508, 2002.

[2] Ming-Syan Chen, Jiawei Han, Yu, P.S, “Data mining: an overview from a database perspective,” Transactions on Knowledge and Data Engineering, Vol. 8, No.6, pp. 866 – 883, 2002.

[3] 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, 1984.

[4] Christoph Bock, Thomas Lengauer, “Computational epigenetics,” Bioinformatics, Vol. 24, No.1, pp. 1-10, 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, 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, 1996.

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

[8] Breiman L, Friedman J, Olshen R, Stone C, "Classification and Regression Trees," Wadsworth International Group, 1984.

[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, 1998.

[10] Niels Landwehr, Mark Hall, Eibe Frank, “Logistic Model Trees,” Machine Learning, Vol. 59, No. 1-2, pp. 161-205, 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, 1992.

[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, 1984.

[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, 1999.

[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, 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 tumours: a gene expression study," Lancet, Vol. 359, No. 9314, pp. 1301-1307, 2002.

[17] Fort G, Lambert S, "Classification using partial least squares with penalized logistic regression", Bioinformatics, Vol. 21, No. 7, pp. 1104-1111, 2005.

[18] D.S.V.G.K.Kaladhar P.V. Nageswara Rao, B.L.V. Ramesh Naidu Rajana, "Confusion matrix analysis for evaluation of speech on Parkinson disease using WEKA and Matlab," IJEST, Vol. 2, No.7, pp. 2734-2737, 2010.

[19] Forgionne G, Gangopadhyay A, Adya M, "Cancer Surveillance Using Data Warehouse, Data Mining, and Decision Support Systems," In Top Health Inf Manage, Vol. 21, No. 1, pp. 21-34, 2000.

[20] http://lib.stat.cmu.edu/DASL/Datafiles/CancerSurvival.html.

[21] Rosemary Yancik, Lynn A Ries, "Cancer in older persons. Magnitude of the problem-how do we apply what we know?," cancer, Vol. 74, Issue Supplement S7, pp. 1995–2003, 1994.

[22] Andrea L Houston, Hsinchun Chen, Susan M Hubbard, Bruce R Schatz, Tobun D Ng, Robin R Sewell, Kristin M Tolle, "Medical Data Mining on the Internet: Research on a Cancer Information System," Artificial Intelligence Review, Vol.13, No. 5-6, pp. 437-466, 1999.

[23] Cios KJ, Moore GW, "Uniqueness of medical data mining," Artificial Intelligence in Medicine, Vol. 26, No. 1, pp. 1-24, 2002.

[24] Eibe Frank, Mark Hall, Len Trigg, Geoffrey Holmes, Ian H. Witten, "Data mining in bioinformatics using Weka," Bioinformatics, Vol. 20, No. 15, pp. 2479-2481, 2004.