Design of Decision Support System in the Metastatic Colorectal Cancer Data Set and Its Application

T. Pala and A.Y. Camurcu

***Abstract*— Developing the system which will help doctors with the result to be obtained from the medical data sets by realizing the design of the medical decision support system in which data mining methods are used is the primary objective of this study.**

The survivability condition of metastatic colorectal cancer disease has been predicted using data mining methods in the developed system. In the process of data mining, after the phase of data preprocessing, Support Vector Machines, Naive Bayes, Decision Trees, Artificial Neural Networks, Multilayer Perceptron, Logistic Regression algorithms have been used.

In the study, two different medical decision support system models, classifying prediction and hybrid models, have been developed, and the results obtained from the two models have been compared after being examined. While the most successful algorithm is Support Vector Machines in Classifying Prediction Model in which only classification algorithms are applied, it is Decision Trees and Artificial Neural Networks which are the most successful algorithms with an accuracy rate of 100 per cent in Hybrid Prediction Model. In consequence of the classifying processes, when the accuracy rates of the models are examined, it is seen that while the accuracy rate of Classifying Prediction Model is 65-70%, this rate reaches 95-100% in Hybrid Prediction Model. It has been seen that accuracy rates have elicited very high and very close values for all of the algorithms with the realized hybrid structure, that is, with clustering, in the applications of the classifying algorithms combined.

Index Terms— ***Medical Decision Support System, Data Mining, Classification, K-Means Clustering, Metastatic Colorectal Cancer Disease Data Set***

I. INTRODUCTION

THE large intestine that forms some of the digestive system is composed of two parts named as colon and rectum. Colon is the name for the most of the large intestine; rectum is the about 10-cm part before anus. Cancer cells can spread out of the large intestine after colorectal cancer is formed, and can make the other organs metastasis.

T. PALA, works in University of Duzce, Duzce, Turkey, (e-mail: palatuba@gmail.com , tubapala@duzce.edu.tr).

A.Y. CAMURCU, is with Department of Computer Engineering University of Fatih Sultan Mehmet Vakif University, Istanbul, Turkey, (e-mail: ycamurcu@fsm.edu.tr).

The cancers that appear in the large intestine (colon + rectum) are named as colorectal cancers. Every year, colorectal cancer develops in over 1 million people in the World. Mortality rate is around 33% [1]. In the case of that the cancer cells spread out of the tissue from which they take their roots to another tissue, the tumor cells that are formed in the new tissue show the former tissue cells properties and are named identically. For example; the colorectal cancer cells which make the liver metastasis are actually the large intestine cancer cells settled in the liver tissue. In this case, the disease is named as "metastatic colorectal cancer", not "liver cancer", and the treatment of these cancer cells is done considering not liver cancer but colorectal cancer [1].

According to Bijan and Safaee, although survivability length has increased in recent years, mortality rate is still high. Studies done in this field have determined a couple of factors for patients' survivability prediction; however, survivability length hasn't increased significantly. Among these factors, those which are related to cancer early diagnosis are taken as the most important ones [2].

Grumett and his friends have done the study of survival prediction using artificial neural networks and logistic regression analysis on colorectal cancer patients' data. In this study, data that belong to 403 patients are analyzed. Results obtained from the logistic regression analysis and artificial neural network models have been evaluated concerning characteristically drawings, and the compatibility of the models have been compared. Logistic regression model has resulted in 66%, artificial neural network model in about 78%. It has been seen that neural network model is superior to logistic regression analysis model according to these results [3].

Biglarian and his friends collected the data of this study in which they aimed to make distant metastasis prediction on colorectal cancer patients using artificial neural networks as 1219 records from a hospital in Iran between the years 2002 and 2007. In the metastasis prediction, artificial neural networks and logistic regression models C_index (concordance _index) and AUROC (Area Under Receive Operating Characteristic) parameters have been compared. In the application which has been done using R2.14.1 statistical software, C_index has obtained ANN 0.82 Logistic Regression 0.77 results for ANN 0.812 Logistic Regression 0.779 AUROC. These results have demonstrated that the prediction accuracy of ANN is better than Logistic

Regression. This study shows that ANN is an appropriate method which can be used in metastasis prediction [4].
Shenbaga and Vijayalakshmi have searched the subtraction of survival rate with the help of feedback networks from the SEER colorectal cancer data. When Multilayer Perceptron and Feedback networks are compared through classical statistical algorithms, the use of data mining algorithms have been seen to be acceptable in survival prediction [5].

Fathy has analyzed the data of over 100000 colorectal cancer patients taken from SEER through artificial neural networks for colorectal cancer survival prediction. In SEER data set, there are over 70 inputs. Out of these inputs, 20 inputs which are not related to other variables have been selected. The best prediction accuracy rate has been sought to be reached with the most appropriate ANN architecture. For survival prediction, two models have been formed. In the first model, the experimental results have demonstrated that the highest prediction rate is 84.73% with 19 inputs. In the second model, it has been sought to reach the highest accuracy value with the smallest number of inputs. The number of inputs has been decreased to 14, and the number of latent neurons from 11 to 8. As a result of these measurements, it has been seen that 73.68% of the number of the inputs and 72.72% of the number of the latent neurons are sufficient to form an appropriate ANN architecture [6].

In this study, cancer data collected by the Oncology Department at Erciyes University Medical Faculty. These data are the data of 200 patients from 27 fields, and they include demographic information and biochemical and pathological results of the patients. The rapid and substantial technological development has enabled database systems to develop, too. Developments in database systems have enabled collecting data, creating databases, managing data and data to be transferred to electronic environment more easily and to be stored more safely and cheaply. Developments in database systems have increased the use of data mining. The purpose of data mining is to analyze the data collected in databases examining them with mathematical and statistical methods, and to reveal the available rules, relationships, structures or different unfamiliar information. The application has been done using RapidMiner data mining program. RapidMiner has been developed using Java code by Yale University. For categorization and regression, there are many characteristics such as a variety of algorithms, decision tress, Bayesian, logical clusters, association rules and clustering algorithms, data preprocessing, normalizing, filtering properties, genetic algorithm, artificial neural networks, data analysis in 3D. RapidMiner can import data from the databases Oracle, Microsoft SQL Server, PostgreSQL, MySQL [7]. Two models have been designed: Classifying Prediction Model and Hybrid Prediction Model. Predicting success rate of the models have been evaluated by applying the designed models on the data set, and their effects on the predicting results have been observed.

II. Data Mining Models

A. Data Mining Classification Models

Human mentality tends to categorize and classify the objects, events, situations around. Thus, he can understand and talk about the objects and events better. Categorization process in data mining is the process of categorizing the available data according to the determined features and of predicting this data category when some new data are added.

Decision Tree: Decision Tree is one of the most frequently used classification models to analyze data. A decision tree is composed of root, trunk and leaf joints. Considering that the structure of a tree develops from root to leaves, it's formed from top to bottom. The most outer joint is the root joint. Each inner joint of the tree is separated to make the best decision with the help of algorithms [8]. Tree leaves form the category tags, namely, categorical characteristics, making the data in the data set groups.

Artificial Neural Networks: Artificial Neural Networks (ANN) normally enable people to learn from similar or different events they experience, face, observe, to get new knowledge, to generalize those events, relating them with each other, to learn from their mistakes if they make any, and to make decision using all of these in an event they come across. ANN is inspired from human's problem solving with the abilities of thinking, observing, learning from mistakes, trial-error, that is, in a more general speaking, learning. An artificial neural network is an information processing system based on human cognition simulation. It is composed of plenty of calculating neural units attached together. These units are called neurons. The neurons in the nervous system form the network getting attached. ANN is a model developed on layers.

The neurons in the network are arranged all along the layers. A neuron in a layer is connected to all of the neurons in the next layer. Every neuron in the layers takes informative signals from all of the neurons in the former layer, and multiplies every informative signal with their massive values. Process is done with the activation function to get the output, collecting weighted inputs. The output of this function is transmitted to all of the neurons in the next layer. This process is completed after done by the neuron in the output layer, too [9].

Multilayer Perceptron: MLP (Multilayer Perceptron) is an artificial neural network model which is mostly used and learns best [10]. It is known that Multilayer Perceptron has a very strong function in classifying prediction problems [10]. The purpose of this model is to minimize the difference between the target result (output) of the network and the attained result. This model is expressed as propagation algorithm since it makes the mistake spreading it over the network, or as backpropagation since it uses backpropagation learning algorithm in the process of that the multilayer perceptron is getting trained.

Naive Bayes: Bayes classifiers are statistical classifiers. Bayes predicts the membership probabilities of the data, that is, their probability about belonging to a specific category. Bayes classifier is based on the Bayes theorem explained below:

A sample in a data set is composed of the input values X = {x1,x2,xm}. If it is pretended that the total number of the categories is m, the probability calculations are done with Equation 1 for the sample whose category is to be determined [11].

$$P(Ci/X)\frac{P(X/Ci)P(Ci)}{P(X)} \quad (1)$$

p(x|Ci) : the probability for a sample from Category *i* to be *x*
P(Ci) : the first probability of Category *i*
p(x) : the probability for any sample to be *x*
P(Ci|x) : the probability for sample x to be from Category i (last probability)

Support Vector Machines: Support vector machines is a popular method for binary categorization. SVMs can be seen as a perceptron which tries to find a hyper plane that separates the data. The perceptron simply tries to find a hyper plane. It does this process in a way that it separates the data of the hyper plane best. However, it is preferred for the hyper plane to separate the categories as much as possible because it is sought that the hyper plane generalizes the invisible data best with this way of separation. The structure that technically best calculates the separation of data is geometric boundaries. The distance of the hyper plane means the largest boundary which will separate the closest points in the data set, saying in general, means the largest boundary which will categorize the data best [12]. It is not enough for the data in the data set to be in the right side, in the right category of the hyper plane in SVM. Also, they should be in some distance to the hyper plane for a good generalization. Determination of the most appropriate hyper plane directly affects the success of the categorization.

Logistic Regression: The statistical analysis method which is used to express the relationship between a dependent variable and one or more than one independent variables numerically is called Regression Analysis. The purpose in Regression Analysis is to calculate to what extent the independent variables affect the dependent variable, that is, to predict the value of the dependent variable starting out of predicted variables. Logistic Regression is primarily used to predict binary or multi-category dependent variable categories; because response variable is separate, and cannot be directly modeled with linear regression. Therefore, instead of prediction of the point, prediction model of the event occurring is built. In case of the ratio to be more than 50% in the binary categorization problem, category '1' is meant to be assigned to the category determined as category '0' for other cases [13]. The purpose of Logistic Regression Analysis is to calculate to what extent independent variables affect dependent variables.

B. Data Mining Clustering Model

K-Means: It is the best and mostly used clustering algorithm. It is used to divide into k number of categories which have been determined according to quality and characteristics of the records in the data set beforehand. Categorization is done by locating the records in the data set in the closest cluster centers. The success of the clustering process is determined by intra-cluster similarity to be maximum, inter-cluster to be minimum. In other words, objects in the cluster are to be located in the closest way, clusters in the most distant way.

III. APPLICATION

A. Data Set Preprocessing Procedure

In the quality of data analysis results, data quality is very important. Data should be meticulously selected, analyzed, synthesized and prepared for data mining analysis. In this study, quality values have been placed into certain intervals with their mean and standard deviation values by applying data transformation with Z-score normalization in the Metastatic Colorectal Cancer data set. For the qualities in the cancer data set to be equal, z-sore normalization equation has been applied like this:

$$NewQualityValue(i)\frac{QualityValue(i)-Mean(i)}{S\tan dard\ Deviation(i)} \quad (2)$$

B. Design of the Models Applied on the Data Set

In the study, to predict the survival condition of Metastatic Colorectal Cancer patients, two models, Classifying Prediction Model and Hybrid Prediction Model, have been designed. Prediction success ratio has been evaluated by applying the designed models on the data set and the effects of the models on predicting results have been observed.

While applying algorithms on the data set, "10-times cross correction" method has been used as testing method. With this method, the data set is divided into ten parts, and every part is used as test data for one time, as training data for the other nine times. When evaluating the success of the created models, accuracy value has been used.

C. Classifying Prediction Model

Block diagram belonging to the first model, Classifying Prediction Model as we call it, is seen in Figure 1.

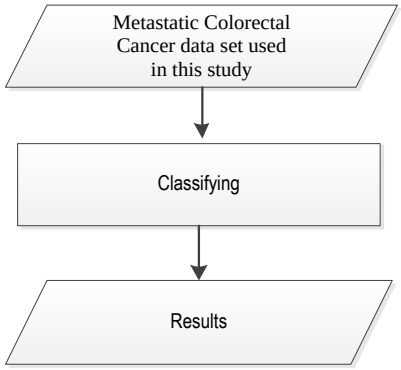

Fig.1. Block Diagram of Classifying Prediction Model

In Classifying prediction model, Metastatic Colorectal Data Set which has been ready for data analysis passing through pre-processing is categorized with selected categorization algorithms. Later, categorization results are compared according to accuracy percentage value which is from the algorithm evaluation criteria. Classifying algorithms have been applied in the data set, and their accuracy values have been compared in Table 1.

TABLE I
ACCURACY VALUES OF CLASSIFYING PREDICTION MODEL

Classifying Algorithms	Accuracy (%)
ANN	62,50
Naive Bayes	64,00
Decision Trees	64,00
MLP	65,00
LR	66,00
DVM	71,00

In Table 1, algorithms and accuracy percentages are seen. When the values are examined, it is seen that the algorithm of artificial neural network has an accuracy percentage of 62,50%, Bayes and decision trees algorithms 64%. MLP with an accuracy ratio of 65%, Logistic Regression with 66%, have very close values to other algorithms. On these data, the algorithm with maximum accuracy ratio is SVM algorithm with 71%.

D. Hybrid Prediction Model

The model applied in Hybrid Prediction Model is for K-Means clustering algorithm to be applied on the data set which has been pre-processed, and for the classifying algorithms of the data (patient records) whose categories have been mispredicted as a result of clustering to be tested on the new data set by clearing the data. Thus, Hybrid Prediction Model is a hybrid model which is formed with clustering+classifying techniques. In Figure 2, the block diagram of Hybrid Prediction Model is seen.

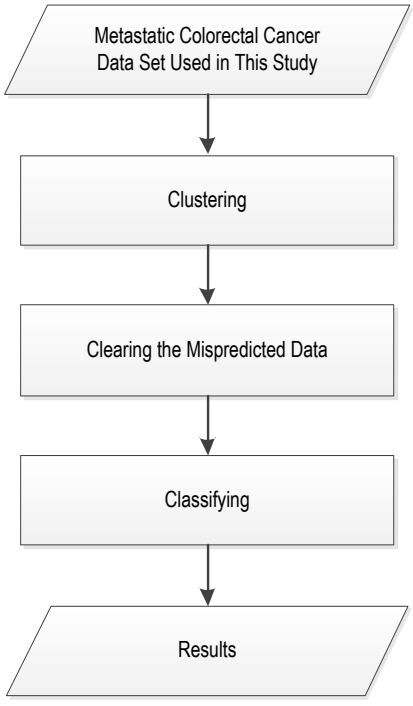

Fig.2. Block Diagram of Hybrid Prediction Model When K-Means clustering algorithm is applied on the data set, emerging results are seen.

TABLE II
K-MEANS CLUSTERING ALGORITHMS CLUSTERING CATEGORIES

Cluster attribute (clusters)	Number of the samples	Number of the data whose categories are mispredicted
Cluster 1 → alive Cluster 0 → dead	200	90

90 misclassified data seen in Table 2 are deleted, and when classifying algorithms are applied on the new data set, the results are as below:

TABLE III
ACCURACY VALUES OF HYBRID PREDICTION MODEL

Classifying Algorithms	Accuracy (%)
LR	96,36
Naive Bayes	97,27
MLP	98,18
SVM	99,09
ANN	100,00
Decision Trees	100,00

Accuracy values of classifying of this model give pretty good results. The algorithm which has the lowest accuracy value is Logistic Regression algorithm with 96,36%. Bayes has an accuracy percentage of 97,27%, MLP 98,18%. It is seen that artificial neural networks and decision trees are the best algorithms with a 100% classifying success in Hybrid Prediction Model. SVM is the second best algorithm with a high success percentage of 99,09%.

IV. RESULTS AND SUGGESTIONS

In this study, survival conditions of Metastatic Colorectal Cancer patients have been predicted by benefitting from data mining algorithms. Predicting process has been done by classifying the data with RapidMiner. In RapidMiner, also the mispredicted data have been detected.

For predicting survival condition, two different models, Classifying Prediction Model and Hybrid Prediction Model, have been created.

In conducted studies, processes have been done on the data composed of 20 qualities of 200 Metastatic Colorectal Cancer patients. Pre-processing, classifying and clustering processes have been applied on the data. The data have been classified according to 6 classifying algorithms which give the best results.

Decision trees and artificial neural networks have been the most successful algorithms with an accuracy percentage of 100% in Hybrid Prediction Model, while the best algorithm in the classifying model in which only classifying algorithms are applied is SVM.

When the accuracy values of the models are examined as a result of the classifying processes, while accuracy percentages of Classifying Prediction Model are 65-70%, this ratio is seen to have reached 95-100% in Hybrid Prediction Model. With the realized hybrid structure, that is, with clustering, accuracy

percentages have been seen to reveal very high and very close values in applying classifying algorithms by combining them. In predicting survival condition, a very successful predicting model with very high success percentages has been realized in the model formed using clustering and classifying algorithms. New hybrids models can be developed by changing the algorithms used here. On future dates, better results than data mining can be obtained by increasing the number of the data used in this study and creating more extensive patterns. In the future, this application can be developed on other cancer types, too.

Doctors can be provided convenience in the progress of the disease and treatment for patients whose death is predicted by making survival or death prediction of colorectal cancer disease which is seen on many people in the world and in our country beforehand. Also, survival length of patients can be extended or their treatment can be procured by trying some other new treatment methods and taking other precautions.

ACKNOWLEDGEMENT

I would like to thank Assist. Prof. Dr. Veli Berk, academic member in Oncology Department at Erciyes University Medical Faculty, for providing the data of this study, and Assist. Prof. Dr. Çiğdem Pala, academic member in Hematology Department at Erciyes University Medical Faculty, for her help in medical matters.

REFERENCES

[1] http://kolonkanseri.blogspot.com/p/kolorektal-kanser-nedir.html: Kolorektal Kanser Nedir ? (Son erişim: 06.10.2012).

[2] B. Moghimi-Dehkordi, A. Safaee, “An overview of colorectal cancer survival rates and prognosis in Asia”, World J Gastrointest Oncol Vol.4, No.4, pp.71-75, 2012.

[3] S. Grumett, P. Snow, and D. Kerr”, Neural Networks in the Prediction of Survival in Patients with Colorectal Cancer”, Clinical Colorectal Cancer, Vol.2, No.4, 239-244, 2003.

[4] A. Biglarian, E. Bakhshi, M.R. Gohari, R. Khodabakhshi, “Artificial Neural Network for Prediction of Distant Metastasis in Colorectal Cancer”, Asian Pacific Journal of Cancer Prevention, Vol.13, pp.927-930, 2012.

[5] S. Shenbaga Ezhil, and C. Vijayalakshmi, “Prediction Of Colon-Rectum Cancer Survivability Using Artificial Neural Network”, International Journal of Computer Engineering and Technology (IJCET), Vol.3, No.1, January- June 2012.

[6] S.K. Fathy, “A Predication Survival Model for Colorectal Cancer”, proceeding AMERICAN-MATH’11/CEA’11 Proceedings of the 2011 American conference on applied mathematics and the 5th WSEAS international conference on Computer engineering and applications, pp.36-42, 2011.

[7] M. Dener, M. Dorterler, and A. Orman, "Açık Kaynak Kodlu Veri Madenciliği Programları: WEKA’da Örnek Uygulama", Akademik Bilişim’09 - XI. Akademik Bilişim Konferansı, Harran University, 11-13 February 2009 Şanlıurfa. (in Turkish).

[8] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, California, 1993.

[9] C. Beh Boon, M.Z.M. Jafri, S. Lim Hwee, "Mangrove Mapping in Penang Island by Using Artificial Neural Network Technique", Open Systems (ICOS), 2011 IEEE Conference on, pp.245-249, 2011.

[10] D. Delen, G. Walker, A. Kadam, “Predicting breast cancer survivability: a comparison of three data mining methods”, Artificial Intelligence in Medicine, Vol 34, pp.113-127, June 2005.

[11] R. Jiangtao, L. Sau Dan, C. Xianlu, K. Ben, R. Cheng, and D. Cheung, "Naive Bayes Classification of Uncertain Data", Data Mining, 2009. ICDM '09. Ninth IEEE International Conference on, pp.944-949.

[12] A. Menon Krishna, “Large-Scale Support Vector Machines: Algorithms and Theory”, Research Exam, University of California, San Diego, 2009.

[13] H. Arslan, “Web Mining and Data Analysis of a Web Site, Sakarya University, Master Thesis, Sakarya, Turkey, 2008. (in Turkish).

BIOGRAPHIES

TUBA PALA was born in Konya, Turkey. She received the B.S. and M.S degrees from the University of Marmara, Istanbul, Turkey. Since 2011, she is a Lecturer with the Duzce University. Her research interests are in Data Mining and Biomedical Data Processing, Artificial Intelligence.

.

ALI YILMAZ CAMURCU has been a Professor with the Computer Engineering Department, University of Fatih Sultan Mehmet Vakif University, Istanbul, Turkey. He received his Ph.D. from University of Marmara, Istanbul, Turkey in 1996. His research interests include: Computer Architecture, Data Mining, Artificial Intelligence