Middle-East Journal of Scientific Research 24 (12): 3833-3837, 2016
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2016.3833.3837

A Comparative Study on the Performance of Classifiers in Prediction of Rare Clinical Ailments: A Case-Study with Mesothelioma

Krithika Narayanan, Jeevana Chaitra Singumahanti and Shomona Gracia Jacob

Department of Computer Science, SSN College of Engineering, India

Abstract: Data Mining is the analysis of voluminous data in search of relationships which were previously not known. Malignant Mesothelioma (MM) is a very aggressive tumor of the pleura. The objective discussed in this research is to compare the performance of the classifiers in predicting Mesothelioma. This is made possible by comparing the classification algorithms that are traditionally known to perform well such as SMO, J48, Random Forest and Bayes Net. The results of this work report that Random Forest performs best when there are several features contributing to model construction with up to 74% accuracy and J48 performs best when the number of features are significantly less with up to 71% accuracy. Thus, we conclude that when the features are more in number Random Forest performs best and also that we need to build better classifiers when the features which contribute to model construction are very less.

Key words: Data Mining · Mesothelioma · SMO · J48 · Random Forest · Bayes Net · Classification

INTRODUCTION

Data Mining is the process of extracting meaningful patterns from large datasets. This research work aims at studying the effects of data mining techniques in predicting the presence of rare medical conditions such as mesothelioma. It is a cancer that occurs in the mesothelium, a thin membrane encompassing the body's internal organs and cavities. Malignant Mesothelioma is the most serious of all asbestos-related diseases. Exposure to asbestos is the primary cause and risk factor for mesothelioma. It generally results from occupational asbestos exposure, but there are instances of environmental exposure that can also cause the disease [1]. Making a correct diagnosis is particularly difficult as the disease often presents with symptoms that mimic other common ailments. Imaging techniques have been used in the past to diagnose mesothelioma. However a biopsy is required to confirm the disease. Mesothelioma has a poor prognosis [2]. The median survival time for pleural mesothelioma is 12 months from diagnosis. Women, young people, people with low-stage cancers and people with epithelioid cancers have better prognoses. Negative prognostic factors include sarcomatoid or biphasic histology, high platelet counts (above 400,000), age over 50 years, white blood cell counts above 15.5, low glucose levels in the pleural fluid, low albumin levels and high fibrinogen levels. Several markers are under investigation as prognostic factors, including nuclear grade and serum C - reactive protein [3]. Long-term survival is rare. Using data mining techniques to help predict the ailment would be helpful in timely treatment.

In the study presented in this paper, the performance of various classification algorithms in predicting the presence of Mesothelioma from the data is investigated. WEKA (Waikato Environment for Knowledge Analysis) is used to conduct the experiments. It is a non-commercial and open-source data mining system with tools for data pre-processing, classification, regression, clustering, association rules and visualization. Four different classifiers are selected and evaluated, which are Bayesian Networks, SMO, J48 and Random Forest, respectively[4].

MATERIALS AND METHODS

Dataset Description: The data set used in this study, was obtained from Dicle University Faculty of Medicine in Turkey. The data set include 324 Mesothelioma patient data samples and all samples have 34 features. The data is unevenly distributed since it comprises of 228 instances of healthy patients and only 96 cases with Mesothelioma.

Corresponding Author: Krithika Narayanan, Department of Computer Science, SSN College of Engineering, India.

Table 2.2.1: Table indicating details about the dataset

Data set Characteristics:	Attribute Characteristics:	Associated Tasks:	Number of instances:	Number of attributes:	Missing Values:
Multivariate	Real	Classification	324	34	N/A

Data Set Information: Malignant mesotheliomas (MM) are very aggressive tumors of the pleura. These tumors are connected to asbestos exposure. However it may also be related to previous simian virus 40 (SV40) infections and quite possible for genetic predisposition. Molecular mechanisms can also be implicated in the development of Mesothelioma. Rural living is associated with the development of mesothelioma. Soil mixtures containing asbestos, known as white-soil or corak can be found in Anatolia, Turkey and Luto in Greece.

Attribute Information:

The Attributes (Features) Contributing to the Disease Are as Follows: Age, gender, city, asbestos exposure, type of MM, duration of asbestos exposure, diagnosis method, keep side, cytology, duration of symptoms, dyspnoea, ache on chest, weakness, habit of cigarette, performance status, White Blood cell count (WBC), hemoglobin (HGB), platelet count (PLT), sedimentation, blood lactic dehydrogenise (LDH), Alkaline phosphatise (ALP), total protein, albumin, glucose, pleural lactic dehydrogenise, pleural protein, pleural albumin, pleural glucose, dead or not, pleural effusion, pleural thickness on tomography, pleural level of acidity (pH), C-reactive protein (CRP), class of diagnosis.

Feature Selection: Performance of the classifiers can be improved by selecting those features which contribute to the enhancing the accuracy of the classifier [5]. Feature selection can be done in three ways, filter, wrapper and hybrid (both filter and wrapper). Filters are where the attributes are ranked and chosen independent of the classification algorithm used, but with wrappers on the other hand the classification algorithm is taken into account while choosing the features.

In Correlation-based Feature Subset Selection [6], useful feature subsets are those that contain features which help predict the class but are not correlated with another feature. CFS computes a heuristic measure of the "merit" of a feature subset from pair-wise feature correlations and a formula adapted from test theory. Heuristic search is used to traverse the space of feature subsets in reasonable time; the subset with the highest merit found during the search is reported. CFS initially discretizes all continuous features in the training data in order to mete out uniform treatment. But chances of an irrelevant attribute being preferred to a predictive one gets greater with fewer training examples. The selected features are tabulated below.

Table 2.4.1: Table indicating the features selected

Features Selected
Diagnosis method
Keep side
Platelet count (PLT)

Classification: Classification is the technique which determines the class to which the data record belongs to. Classification Algorithms build models from the training data records (data records in which class labels are known) given to it and this model is used to assign a class label to the new data. A good classifier is one which can generalize beyond the training data and correctly classify the new data presented to it [7]. Generally well-performing classifier types are Naïve Bayes, Logistic Regression, Decision trees and SVMs. For this study, we have chosen an algorithm from each category [8].

J48: It is an open source java implementation of C4.5 for Weka, a data mining tool developed by University of Wakaito. This algorithm is an optimized implementation of C4.5 and outputs a decision tree. Decision Trees are tools that use divide-and-conquer strategies as a form of learning by induction [9]. It contains a root node, several intermediate nodes and leaf nodes. Each node contains a decision and the decision leads to classification. Splitting criterion identifies the best node to split upon at the level of the tree [10].

Algorithm [11]:

INPUT:
D //Training data
OUTPUT:
T //Decision tree
DTBUILD (*D)
{
T=φ;
T= Create root node and label with splitting attribute;
T= Add arc to root node for each split predicate and label;

For each arc do
D= Database created by applying splitting predicate to D;
If (stopping point reached for this path) then
T'= create leaf node and label with appropriate class;
Else
T'= DTBUILD(D);
T= add T' to arc;
}

Once the tree is built, it is applied to each instance of training the dataset and classification is done. The tree models the classification process.

Random Forest: Random Forest algorithm builds a forest (collection) of decision trees.

D= { hk (x, Tk) } where k=1,2,3…..L
L- No of decision trees
Tk - Training set built at random and identically distributed.
hk - Tree built from vector Tk and produces output x.

Trees in a Random Forest are built randomly by selecting m (value fixed for all nodes) attributes in each node of the tree; where the best attribute is chosen to divide the node[12]. The selection of a random subset of features is a type of the random subspace method, which, is a way to implement the stochastic discrimination approach to classification. The vector used for training each tree is obtained using random selection of the instances [13]. In Random Forest, to determine the class of an instance, all of the trees indicate an output x (each it's own), where the most voted is selected as the final result. The classification error depends on the strength of individual trees of the forest and the correlation between any two trees in the forest.

SMO: SMO is an improved training algorithm for support vector machines (SVM) [14]. Like other training algorithms, SMO breaks down a large Quadratic Programming (QP) problem into a series of smaller QP problems. Unlike other algorithms, SMO utilizes the smallest possible QP problems, which are solved quickly and analytically, generally improving its scaling and computation time significantly [15].

Bayesian Network: Bayesian networks (BNs) (also known as belief networks) belong to the family of probabilistic graphical models (GMs) [16]. These graphical structures are used to represent knowledge about an uncertain domain. In particular, each node in the graph represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. These conditional dependencies in the graph are often estimated by using known statistical and computational methods. Hence, BNs combine principles from graph theory, probability theory, computer science and statistics.

Formally, A Bayesian network B is an annotated acyclic graph that represents a Joint Probability Distribution over a set of random variables V [17]. The network is defined by a pair $B = <G, \Theta>$ where G is the

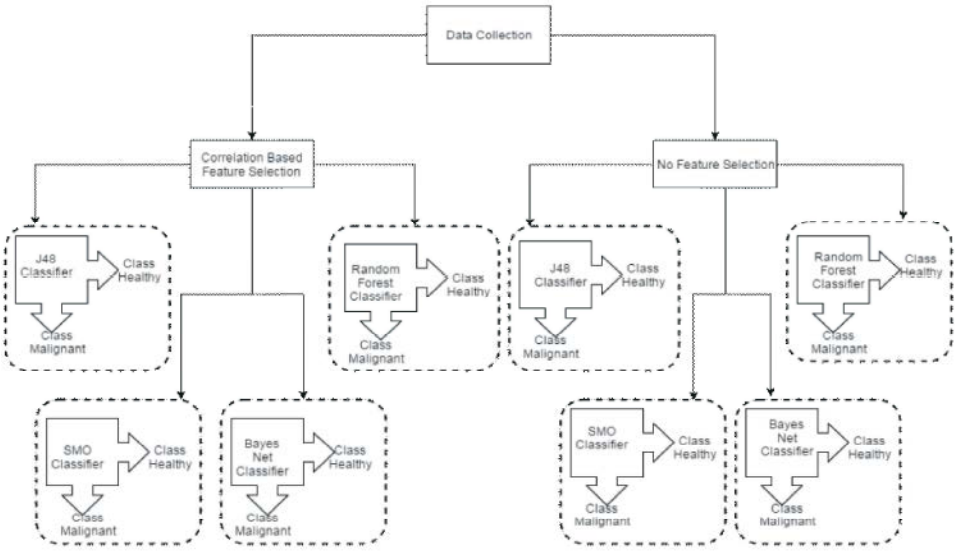

Fig. 1: Proposed methodology for Investigation of Classifier Performance on Mesothelioma Data

DAG whose nodes X_1, X_2,...., X_n represents random variables and whose edges represent the direct dependencies between these variables. The graph G encodes independence assumptions, by which each variable Xi is independent of its non-descendents given its parents in G. The second component Θ denotes the set of parameters of the network. This set contains the parameter $\theta x_i |\pi_i = PB(x_i|\pi_i)$ for each realization x_i of X_i conditioned on π_i, the set of parents of X_i in G.

RESULTS AND DISCUSSIONS

The experimental results are portrayed in two sections. Primary investigations focused on identifying the performance of the classifiers when all the features were included as part of the dataset [18]. Following this, feature selection was applied to remove the irrelevant features and use only the subset of the original feature set for classification.

The results are tabulated in Table 4.1 and Table 4.2 respectively.

Before Feature Selection:

Table 4.1: Table displaying the values found before feature selection

Algorithms Implemented	Time Taken (Sec)	Correctly classified instances	Incorrectly classified instances	Accuracy (%)
Bayes Net	0.18	225	99	69.4444
SMO	0.28	231	93	71.2963
J48	0.04	225	99	69.4444
Random Forest	0.63	242	82	74.6914

The results obtained are including the time taken for computation and the accuracy obtained by 10-fold cross validation.

After Feature Selection:

Table 4.2: Table displaying the values found after feature selection

Algorithms Implemented	Time Taken (Sec)	Correctly classified instances	Incorrectly classified instances	Accuracy (%)
Bayes Net	0.04	225	99	69.4444
SMO	0.01	230	94	70.9877
J48	0	233	91	71.9136
Random Forest	0.2	200	124	61.7284

CONCLUSION

Data mining is the process of analyzing data from many different dimensions or angles, categorizing and summarizing the relationships identified [19]. This paper has investigated the performance of classifiers in predicting the presence of malignant tissue. From the tables above, we see that the accuracy is highest for random forest when all the features have been taken into consideration for model construction, but falls when the numbers of features become lower [20].

As Bayes Net performs consistently, we infer that the number of features used for model construction is irrelevant to its performance. Hence, the results of this work report that Random Forest classifier performs best when there are several features contributing to model construction while J48 performs best with significantly less number of features.

In future, we intend to compare the results of these classifiers on a different dataset and see if similar results are obtained.

REFERENCES

1. Er, O., A.C. Tanrikulu, A. Abakay and F. Temurtas, 2012. An approach based on probabilistic neural network for diagnosis of Mesothelioma's disease. Computers & Electrical Engineering, 38(1): 75-81.
2. Camidge, D.R., D.L. Stockton and M. Bain, 2006. Factors affecting the mesothelioma detection rate within national and international epidemiological studies: insights from Scottish linked cancer registry-mortality data. British journal of cancer, 95(5): 649-652.
3. King, R.D., C. Feng and A. Sutherland, 1995. Statlog: comparison of classification algorithms on large real-world problems. Applied Artificial Intelligence an International Journal, 9(3): 289-333.
4. Jacob, S.G. and R.G. Ramani, 2015. Prediction of Rescue Mutants to Restore Functional Activity of Tumor Protein TP53 through Data Mining Techniques. Journal of Scientific & Industrial Research, 74: 135-140.
5. Li, T., C. Zhang and M. Ogihara, 2004. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20(15): 2429-2437.
6. Ramani, R.G. and S.G. Jacob, 2013. Prediction of P53 mutants (multiple sites) transcriptional activity based on structural (2D&3D) properties. PloS one, 8(2): p.e55401.

7. Jacob, S.G. and R.G. Ramani, 2013. Design and Implementation of a clinical data classifier: A Supervised learning approach. Res. J. Biotech., 8(2): 16-26.
8. Barnaghi, P.M., V.A. Sahzabi and A.A. Bakar, 2012. A comparative study for various methods of classification. In International Conference on Information and Computer Networks, 27(2): 875-81.
9. Khemphila, A. and V. Boonjing, 2010, October. Comparing performances of logistic regression, decision trees and neural networks for classifying heart disease patients. In Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on pp: 193-198. IEEE.
10. Gholap, J., 2012. Performance tuning of J48 Algorithm for prediction of soil fertility. arXiv preprint arXiv:1208.3943.
11. Dunham, M.H., 2006. Data mining: Introductory and advanced topics. Pearson Education India.
12. Díaz-Uriarte, R. and S.A. De Andres, 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7(1): p.1.
13. Guo, L., Y. Ma, B. Cukic and H. Singh, 2004, November. Robust prediction of fault-proneness by random forests. In Software Reliability Engineering, 2004. ISSRE 2004. 15th International Symposium on pp: 417-428. IEEE.
14. Platt, J., 1998. Sequential minimal optimization: A fast algorithm for training support vector machines.
15. Maglogiannis, I., E. Zafiropoulos and I. Anagnostopoulos, 2009. An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers. Applied intelligence, 30(1): 24-36.
16. Geetha Ramani, R. and S. Gracia Jacob, 2013. Prediction of cancer rescue p53 mutants in silico using Naïve Bayes learning methodology. Protein and peptide letters, 20(11): 1280-1291.
17. Muralidharan, V. and V. Sugumaran, 2012. A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis. Applied Soft Computing, 12(8): 2023-2029.
18. Lim, T.S., W.Y. Loh and Y.S. Shih, 2000. A comparison of prediction accuracy, complexity and training time of thirty-three old and new classification algorithms. Machine learning, 40(3): 203-228.
19. Ramani, R.G. and S.G. Jacob, 2013. Improved classification of lung cancer tumors Based on structural and physicochemical properties of proteins using data mining models. PloS one, 8(3): p.e58772.
20. Ramani, R.G. and S.G. Jacob, 2013. Benchmarking classification models for cancer prediction from gene expression data: A novel approach and new findings. Studies Informatics Control, 22(2): 134-143.