Proc. of the Third Intl. Conf. on Advances in Bio-Informatics and Environmental Engineering - ICABEE 2015
Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-078-1 doi: 10.15224/ 978-1-63248-078-1-119

Comparison of Classification Techniques on Dermatological Dataset

[Kemal TUTUNCU, Murat KOKLU]

Abstract— **Data mining is the process of analysing data and summarizing it into useful information. One of main problem in the field of data mining is classification. Having done in this study, Simple Logistic Regression, Bayes Net, Naïve Bayes, Radial Basis Function Network (RBF), Multilayer Perceptron (MLP), Naïve Bayes Tree (NB Tree), Sequential Minimal Optimization (SMO), J48, Random Tree and ZeroR classification methods were applied on dermatology data set by UCI Machine Learning Repository. When comparing the performances of algorithms it's been found that Simple Logistic Regression and Bayes Net have highest accuracies whereas ZeroR had the worst accuracy. The results were also compared with previous studies in the literature. It has been seen that Simple Logistic Regression and Bayes Net had promising results when they compared with the methods used in literature.**

Keywords— ***data mining, classification, j48, bayes net, smo, zeror***

I. Introduction

Data mining has been used by many organizations to extract information or knowledge from large volumes of data and then use the valuable information to make critical business decisions. Consequently, analysis of the collected history data in data warehouse or in data mart can gain better insight into your customers and evaluation of the medical diagnosis and prognosis, improve the quality of decision-making and effectively increase the opportunity of the curability for these vital illness. [1, 2]

Classification in Data Mining is one of the sub-area of data mining. It can be defined as the problem of extracting knowledge from a set of n input examples x1,…,xn characterized by i features a1,…,ai ∈A, including numerical or nominal values, where each instance is labelled with a desired output class label yj ∈C (considering a m class problem C={c1,…,cm}) and the aim is to learn a system capable of predicting this output for a new unseen example in a reasonable way (with good generalization ability) [3]. The system generated by the learning algorithm is a mapping function defined over the patterns Ai→C and it is called a classifier [4].

Having done in this study, some well-known classification techniques were applied to the Dermatology dataset from University of California Irvine Machine Learning Repository (UCI MRL) [5]. The results were compared to the results obtained by previous studies.

The remainder of this paper is organized as follows: In Section 2 previous studies in literature are presented. In Section 3 materials and methods used in this study are described. In Section 4, Simple Logistic Regression, Bayes Net, Naïve Bayes, RBF Network, Multilayer Perceptron, NB Tree, SMO, J48, Random Tree and ZeroR classification methods are applied to Dermatology dataset. The results are compared to each other and the results of the studies in literature. And finally, in Section 5 the conclusions are stated.

II. Literature Survey

As can be seen from the following studies so many classification methods have been applied to Dermatology Dataset by UCI Machine Learning Repository. This study make use of other classification methods to apply same Dermatology Dataset. Thus advantages and disadvantages of the methods used in this study were presented by making comparisons with pioneers.

Tomar and Agarwal (2015), extended the formulation of binary Least Squares Twin Support Vector Machine classifier to multi-class by using the concepts such as ‘‘One-versus-All’’, ‘‘One-versus-One’’, ‘‘All-versus-One’’ and Directed Acyclic Graph (DAG). They applied proposed methods 12 different data set including Dermatology Data set by UCI MRL [6].

Celia et all. (2004) proposed constrained-syntax genetic programming for discovering classification rules by applying the method to the 4 medical datasets [7]. Karaboga and Ozturk (2011) used Artificial Bee Colony (ABC) for data clustering on benchmark problems and the performance of ABC algorithm is compared with Particle Swarm Optimization (PSO) algorithm and other nine classification techniques from the literature. Thirteen of typical test data sets from the UCI MRL are used to demonstrate the results of the techniques. The simulation results indicate that ABC algorithm can efficiently be used for multivariate data clustering [8].

Polat and Gunes (2009) proposed novel hybrid classification system based on C4.5 decision tree classifier and one-against-all approach to classify the multi-class problems including dermatology, image segmentation, and lymphography datasets taken from UCI MRL database [9]. Bahrololoum et all. applied gravitational search algorithm (GSA) to solve clustering and classification problems. The results are compared with ABC and PSO algorithms [10].

Kemal Tutuncu

Department of Electrical – Electronics Engineering, Selcuk University
Konya, TURKEY

Murat Koklu

Department of Computer Engineering, Selcuk University
Konya, TURKEY

Zhang et all. (2009) applied a rough set-based multiple criteria linear programming (RS-MCLP) approach for solving classification problems [11]. Dennis and Muthukrishnan (2014) applied Adaptive Genetic Fuzzy System (AGFS) for medical data classification [12]. Oriol et all. (2008) applied the error correcting output codes (ECOC) technique with SVM [13]. Park ad Kim (2007) applied grey-zone case-based reasoning (GCBR) that makes decisions focusing additional attention on the cases near the cut-off point by interactive communication with users to the medical data sets for classification aims [14]. Tabakhi et all. (2014) presented an unsupervised feature selection method based on ant colony optimization for classification aim. They applied this method to 9 different data sets and obtained promising results [15].

Altıncay and Erenel (2013) proposed to transform the training data of different classes into separate clusters before applying nearest feature line classifier. Spectral clustering based transformation is used for this purpose and it is shown that the accuracies achieved by both the nearest feature line and the shortest feature line segment approach which is the most recent variant of the nearest feature line technique are improved [16].

Senthilnath et all. (2011) applied Firefly Algorithm (FA) for clustering on benchmark problems. They compared the results with other techniques namely ABC, PSO, BayesNet, Multi layer perceptron ANN, RBF, KStar, Ragging, MultiBoast, NBTree, Ridor, VFI [17].

Luukka (2011) applied feature selection method based on fuzzy entropy measures together with similarity classifier. Model was tested with four medical data sets which were, dermatology, Pima-Indian diabetes, breast cancer and Parkinsons data sets [18]. Kim and Choi (2015) proposed a pattern generation method for multi-class classification using logical analysis of data (LAD). Specifically, they applied two decomposition approaches—one versus all, and one versus one - to multi-class classification problems, and develop an efficient iterative genetic algorithm with flexible chromosomes and multiple populations (IGA-FCMP) [19].

Wanga, et all. (2013) addressed the problem of semi-supervised metric learning. They proposed a new regularized semi-supervised metric learning (RSSML) method using local topology and triplet constraints [20]. Bai and Liang (2014) proposed the k-modes type clustering plus between-cluster information for categorical data to apply classification problems [21]. Tzortzis and Likas (2014) used The Min Max k-Means clustering algorithm for classification [22]. All the upper mentioned study is are summarized with the accuracy ratio in Table 1.

TABLE I. THE RESULT OBTAINED IN THE LITERATURE

Algorithm	Accuracy (%)	Referernces
Fuzzy	98,28	Luukka, 2011, [18]
MLP ANN	96,74	Bahrololoum et all., 2012, [10]
MLPANN	96,74	Senthilnath et all., 2011, [17]
Combination of C4.5 decision tree	96,71	Polat and Güneş, 2009, [9]
GP	96,60	Celia et all., 2004, [7]
Bagging	96,53	Bahrololoum et all., 2012, [10]
Bagging	96,53	Senthilnath et all., 2011, [17]
Multiclass SVM	96,52	Oriol et all., 2008, [13]
NFL-SPA	96,32	Altıncay and Erenel, 2013, [16]
SFLS-SFA	96,32	Altıncay and Erenel, 2013, [16]
NFL	96,15	Altıncay and Erenel, 2013, [16]
All-Versus-One-LSTSVM	96,11	Tomar and Agarwal, 2015, [6]
GSA	95,88	Bahrololoum et all., 2012, [10]
ECOC-ONE	95,83	Oriol et all., 2008, [13]
SFLS	95,55	Altıncay and Erenel, 2013, [16]
Kstar	95,34	Bahrololoum et all., 2012, [10]
Kstar	95,34	Senthilnath et all., 2011, [17]
RNFLS	95,27	Altıncay and Erenel, 2013, [16]
ABC	94,57	Karaboga and Ozturk, 2011, [8]
ABC	94,57	Bahrololoum et all., 2012, [10]
ABC	94,57	Senthilnath et all., 2011, [17]
FA	94,57	Senthilnath et all., 2011, [17]
C4.5	94,48	Kim and Choi, 2015, [19]
PSO	94,24	Karaboga and Ozturk, 2011, [8]
PSO	94,24	Senthilnath et all., 2011, [17]
PSO	93,92	Bahrololoum et all., 2012, [10]
One-Versus-All MLSTSVM	93,85	Tomar and Agarwal, 2015, [6]
VFI	92,40	Senthilnath et all., 2011, [17]
VFI	92,40	Bahrololoum et all., 2012, [10]
IGA-FCMP	92,18	Kim and Choi, 2015, [19]
RIDOR	92,08	Senthilnath et all., 2011, [17]
Ridor	92,08	Bahrololoum et all., 2012, [10]
DAG MLSTSVM	91,27	Tomar and Agarwal, 2015, [6]
NB	90,56	Tabakhi et all., 2014, [15]
SVM	90,24	Tabakhi et all., 2014, [15]
One-Versus-One MLSTSVM	89,56	Tomar and Agarwal, 2015, [6]
C4.5	89,10	Celia et all., 2004, [7]
MC-LAD	89,07	Kim and Choi, 2015, [19]
DT	88,56	Tabakhi et all., 2014, [15]
Multi-SVM	87,18	Tomar and Agarwal, 2015, [6]
CN2	87,10	Kim and Choi, 2015, [19]
MBSVM	86,69	Tomar and Agarwal, 2015, [6]
BGP	86,20	Celia et all., 2004, [7]
C4.5 Decision tree classifier	84,48	Polat and Güneş, 2009, [9]
Twin KSVC	84,06	Tomar and Agarwal, 2015, [6]
TABATA	80,66	Kim and Choi, 2015, [19]
RBF	65,34	Bahrololoum et all., 2012, [10]
RBF	65,34	Senthilnath et all., 2011, [17]
MultiBoost	46,74	Bahrololoum et all., 2012, [10]
Multiboost	46,74	Senthilnath et all., 2011, [17]

III. Metarial And Method

A. *Dataset*

The dermatology dataset used in this study was taken from UCI Machine Learning Repository [5]. The aim of the dermatology data set is to diagnose one of six possible types of eryhemato-squamous diseases (6 classes). Twelve clinical and 24 histopathological measurements of the patient are given (36 attributes). The data set contains 366 patterns, which are to be used for both classifier design and testing [23].

B. *Software-WEKA*

Weka (Waikato Environment for Knowledge Analysis) written in Java, developed at the University of Waikato, New Zealand [24]. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All techniques of Weka's software are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported) [25].

C. *Methods*

Simple Logistic Regression: Simple Logistic Regression sometimes called the logistic model or logit model, analyses the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurrence of an event by fitting data to a logistic curve. There are two models of logistic regression, binary logistic regression and multinomial logistic regression [24].

Bayes Net: It is probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases [24].

Naive Bayes: Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers [24].

RBF Network: It is an artificial neural network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. Radial basis function networks have many uses, including function approximation, time series prediction, classification, and system control [24].

Multilayer Perceptron: A multilayer perceptron (MLP) is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable [24].

NB Tree: NB-tree is a tree data structure that keeps data sorted and allows searches, sequential access, insertions, and deletions in logarithmic time. The NB-tree is a generalization of a binary search tree in that a node can have more than two children. Unlike self-balancing binary search trees, the B-tree is optimized for systems that read and write large blocks of data. It is commonly used in databases and file systems [24].

SMO: Sequential Minimal Optimization (SMO) is a new algorithm for training Support Vector Machines (SVMs). Training a support vector machine requires the solution of a very large quadratic programming (QP) optimization problem. SMO breaks this large QP problem into a series of smallest possible QP problems. These small QP problems are solved analytically, which avoids using a time-consuming numerical QP optimization as an inner loop [24].

J48: J48 algorithm of Weka software is a popular machine learning algorithm based upon J.R. Quilan C4.5 algorithm. All data to be examined will be of the categorical type and therefore continuous data will not be examined at this stage. The algorithm will however leave room for adaption to include this capability [24, 25].

Random Tree: In mathematics and computer science, a random tree is a tree or arborescence that is formed by a stochastic process. Types of random trees include: Uniform spanning tree, Random minimal spanning tree, Random binary tree, Random recursive tree, Treap or randomized binary search tree, Brownian tree, Random forest, and Branching process, a model of a population in which each individual has a random number of children [24].

ZeroR: ZeroR is the simplest classification method which relies on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods. It constructs a frequency table for the target and select its most frequent value [24].

IV. Implementation

After applying Simple Logistic Regression, Bayes Net, Naïve Bayes, RBF Network, Multilayer Perceptron, NB Tree, SMO, J48, Random Tree and ZeroR classification methods on Dermatology Datasets by using WEKA, the accuracy of each method was obtained as in Table 2.

TABLE II. ACCURACY OF EACH METHOD USE IN THIS STUDY

Method	Accuracy (%)
Simple Logistic Regression	97,8142
Bayes Net	97,5410
Naive Bayes	97,2678
RBF Network	96,4481
Multilayer Perceptron	96,1749
NB Tree	95,6284
SMO	95,3552
J48	93,9891
Random Tree	87,4317
ZeroR	30,6011

ISBN: 978-1-63248-078-1 doi: 10.15224/ 978-1-63248-078-1-119

As can be seen from Table 2 the highest accuracy ratios are obtained from Simple Logistic Regression, Bayes Net and Naive Bayes. Since both Bayes Net and Naive Bayes based on probability theory and network only Bayes Net that has slightly higher accuracy ratio mentioned in the abstract and conclusion sections of this study. The worst ratio was obtained by ZeroR method.

When the literature is searched for the classification on Dermatology Dataset It has been seen that the highest ratio is obtained by feature selection method based on fuzzy entropy measures that is studied Luukka (18) as 98.28%.

V. Conclusion

Having done in this study the performances of Simple Logistic, Bayes Net, Naïve Bayes, RBF Network, Multilayer Perceptron, NB Tree, SMO, J48, Random Tree and ZeroR methods were evaluated in terms of classification accuracy on dermatology datasets. When comparing the performances of algorithms it's been found that Simple Logistic Regression (97,8142%) and Bayes Net (97,5410%) have highest accuracies whereas ZeroR (30,6011%) had the worst accuracy. The results were also compared with previous studies in the literature. It has been seen that Simple Logistic Regression and Bayes Net had promising results when they compared with the methods used in literature that gives highest result (98.2800%) named feature selection method based on fuzzy entropy measures. For the future work more classification algorithms should be applied to more datasets to see impacts of the different performance of algorithms on different datasets.

Acknowledgment

This study has been supported by Scientific Research Project of Selcuk University.

References

[1] Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), 570–577. July–August.

[2] Zhang Zhiwang, Shi Yong, Gao Guangxia, A rough set-based multiple criteria linear programming approach for the medical diagnosis and prognosis, Expert Systems with Applications 36 (2009) 8932–8937.

[3] Duda R.O., Hart P.E., Stork D.G., Pattern Classification (second ed.), John Wiley ,2001.

[4] Mikel Galar, Joaquín Derrac, Daniel Peralta, Isaac Triguero, Daniel Paternain, Carlos Lopez-Molina, Salvador García, José M. Benítez, Miguel Pagola, Edurne Barrenechea, Humberto Bustince, Francisco Herrera, A survey of fingerprint classification Part I: Taxonomies on feature extraction methods and learning models, Knowledge-Based Systems, Volume 81, June 2015, Pages 76-97.

[5] Blake A.C.L. and Merz C.J., University of California at Irvine Repository of Machine Learning Databases, 1998, http://www.ics.uci.edu/~mlearn/MLRepository.html.

[6] Tomar Divya and Agarwal Sonali, A comparison on multi-class classification methods based on least squares twin support vector machine, Knowledge-Based Systems 81 (2015) 131–147.

[7] Celia C., Bojarczuka, Heitor S. Lopesa, Alex A. Freitasb, Edson L. Michalkiewiczc, A constrained-syntax genetic programming system for discovering classification rules: application to medical data sets, Artificial Intelligence in Medicine 30 (2004) 27–48.

[8] Karaboga Dervis and Ozturk Celal, A novel clustering approach: Artificial Bee Colony (ABC) algorithm Applied Soft Computing 11 (2011) 652–657.

[9] Polat Kemal and Gunes Salih, A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems, Expert Systems with Applications 36 (2009) 1587–1592.

[10] Bahrololoum Abbas, Nezamabadi-pour Hossein, Bahrololoum Hamid, Saeed Masoud, A prototype classifier based on gravitational search algorithm, Applied Soft Computing 12 (2012) 819–825.

[11] Zhiwang Zhang, Yong Shi b,c, Guangxia Gao, A rough set-based multiple criteria linear programming approach for the medical diagnosis and prognosis, Expert Systems with Applications 36 (2009) 8932–8937.

[12] Dennis, B. and Muthukrishnan S., AGFS: Adaptive Genetic Fuzzy System for medical data classification, Applied Soft Computing 25 (2014) 242–252.

[13] Oriol Pujola, Sergio Escalerab, Petia Radevab, An incremental node embedding technique for error correcting output codes Pattern Recognition 41 (2008) 713 – 725.

[14] Park Yoon-Joo, Kim Byung-Chun, An interactive case-based reasoning method considering proximity from the cut-off point Expert Systems with Applications 33 (2007) 903–915.

[15] Tabakhi Sina, Moradin Parham, Akhlaghian Fardin, An unsupervised feature selection algorithm based on ant colony optimization, Engineering Applications of Artificial Intelligence32(2014)112–123.

[16] Altınçay Hakan and Erenel Zafer, Avoiding the interpolation inaccuracy in nearest feature line classifier by spectral feature analysis, Pattern Recognition Letters 34 (2013) 1372–1380.

[17] Senthilnath J., Omkar S.N., Mani V., Clustering using firefly algorithm: Performance study, Swarm and Evolutionary Computation 1 (2011) 164–171.

[18] Luukka Pasi, Feature selection using fuzzy entropy measures with similarity classifier, Expert Systems with Applications 38 (2011) 4600–4607.

[19] Kim Hwang Ho and Choi Jin Young, Pattern generation for multi-class LAD using iterative genetic algorithm with flexible chromosomes and multiple populations, Expert Systems with Applications 42 (2015) 833–843.

[20] Wanga Qianying, Yuen Pong C, Feng Guocan, Semi-supervised metric learning via topology preserving multiple semi-supervised assumptions, Pattern Recognition 46 (2013) 2576–2587.

[21] Bai Liang and Liang Jiye, The k-modes type clustering plus between-cluster information for categorical data, Neurocomputing 133 (2014)111–121.

[22] Grigorios Tzortzis and Aristidis Likas, The Min Max k-Means clustering algorithm, Pattern Recognition 47 (2014) 2505–2516.

[23] Güvenir H Altay, Demiröz Gülsen, Ilter Nilsel, Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals, Artificial Intelligence in Medicine vol 13, pages 147.

[24] WEKA, http://www.cs.waikato.ac.nz/~ml/weka/ Last access: 10.04.2015.

[25] Rohit Arora and Suman, Comparative Analysis of Classification Algorithms on Different Datasets using WEKA, International Journal of Computer Applications (0975 – 8887) Volume 54– No.13, September 2012.