Available online at www.elixirpublishers.com (Elixir International Journal)

Advanced Engineering Informatics

Elixir Adv. Engg. Info. 69 (2014) 23117-23118

ISSN: 2229-712X

A comprehensive scalability analysis of classification accuracies obtained from original and synthetic medical dataset

Rajkumar.N*, Jaganathan.P and Sangeetha T.S

Department of Computer Applications, PSNA College of Engineering and Technology, Dindigul, Tamil nadu, India.

ARTICLE INFO

Article history:

Received: 28 May 2013;

Received in revised form:

29 March 2014;

Accepted: 10 April 2014;

Keywords

Dimensionality reduction,

Thyroid dataset,

Synthetic dataset,

MLP,

C4.5.

ABSTRACT

Medical diagnosis is very much essential for the human beings to survive. Earlier the diagnosis, lesser is the pain. The disease unattended will lead to extreme conditions and greater pain, loss of health and high expenditure. This paper focuses on the comprehensive scalability analysis conducted on two kinds of thyroid dataset: original and synthetic. The results obtained with original dataset consisting of 215 instances with 5 features have been compared with the results obtained with synthetic dataset consisting of 1075 instances with 25 features. Initially the features were reduced to 3 features and 13 features on both the cases respectively. Subsequently classification accuracy is obtained using top most performing classifiers namely C4.5 and Multi layer perceptron. The results obtained are promising and the dimensionality reduction technique has proved better for smaller datasets as well as for larger datasets.

Introduction

Feature selection is an important part of pattern recognition, machine learning and medical diagnosis. A suitable representation of data from all features is an important problem in machine learning and data mining problems. All original features cannot always be beneficial for classification tasks. Some features are irrelevant in distribution of datasets. These features can decrease the classification performance. In order to increase both the classification performance and to reduce computation cost of classifier, the feature selection process should be used in classification problems (Cao, Shen, Sun, Yang, & Chen, 2007).

Thyroid disease is prevalent especially among women in most part of the world. The Thyroid gland is the one of the largest glands among the endocrine glands. The thyroid gland is placed in the middle of the lower neck, below the larynx and above the clavicles (Esin, Akif and Dervi, 2011; Thyroid gland, 2007). The most common thyroid problems are hypothyroidism, hyperthyroidism and thyroid nodules.

This study focuses on the comprehensive scalable analysis conducted on thyroid dataset with smaller dataset and larger dataset. Improved F-Score method proposed by Xie and Wang, 2011 is used for feature selection on both cases. Multilayer Perceptron and C4.5 algorithms are used for comparison of classification accuracy for both smaller and larger dataset. The results obtained have shown better performance.

The rest of the paper is organized as follows. Section 2 discusses review of literature. Section 3 describes the research methodology carried out. Section 4 concentrates on results and discussion. Section 5 gives conclusions and future work.

Review of literature

Several studies have been reported focusing on thyroid disease diagnosis (Ozyılmaz and Yıldırım, 2002; Polat et al., 2007). In 2004, Pasi obtained the following accuracies with Linear Discriminant Analysis (81.34%), C4.5-1 (93.26%), C4.5-2 (92.81%), C4.5-3 (92.94), MLP (96.24%) and DIMLP (94.86%). Polat et al. (2007) obtained 85.00% and 81.00% of accuracies using AIRS with Fuzzy weighted pre-processing and AIRS respectively. Keles and Keles (2008) obtained 95.33% classification accuracy using Neuro Fuzzy Classification (ESTDD with NEFCLASS-J). Feyzullah Termurtas (2009) produced different classification accuracies using MLNN with LM (92.96%) for 3-fold cross validation, PNN (94.43%) for 3-fold cross validation, LVQ (89.79%) for 3-fold cross validation, MLNN with LM (93.19%) for 10-fold cross validation, PNN (94.81%) for 10-fold cross validation, LVQ (90.05%) for 10-fold cross validation. Doaganteken et al. (2010) obtained classification accuracy of 97.67% with ADSTG method. Doganteken et al. (2011) obtained 91.86% of classification accuracy using GDA-WSVM. Jaganathan and Rajkumar obtained classification accuracy of 93.49% (2012).

Research Methodology

Feature Selection is very much important in the field of data mining and medical diagnosis. A suitable representation of data from all features is important to solve the problems in machine learning and data mining. All original features can become beneficial for classification only if it is suitably represented. Some may become irrelevant if it is not done so. These features will become responsible for the decrease in the classification performance. To increase the classification performance and to reduce computation cost, feature selection process should be used in classification problems (Cao, Shen, Sun, Yang, & Chen, 2007).

There are two datasets used in the study. One is the dataset which contains data with 3 classes, 215 samples and 5 features taken from the most commonly used UCI machine learning repository.(Hoshi et al., 2005; Polat et al., 2007; Esin, Akif and Dervi, 2011 <http://archive.ics.uci.edu/ml/>). The three classes are normal, hyper and hypo functions of the thyroid gland. Among the 215 samples, 150 samples belong to normal- function, 35 samples belong to hyper-function and 30 samples belong to hypo-function of the thyroid gland dataset. The five features are T3-resin uptake test (A percentage) (F1), Total serum thyroxin as measured by the isotopic displacement method (F2), Total serum triiodothyronine as measured by radioimmuno assay (F3), Basal thyroid-stimulating hormone (TSH) as measured by radioimmuno

Tele:

E-mail addresses: rknpsna@gmail.com

assay (F4). Maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value (F5). The second is the synthetic dataset which contains data with 3 classes, 1075 samples and 25 features. The three classes are normal, hyper and hypo functions of the thyroid gland. Among the 1075 samples, 450 samples belong to normal- function, 175 samples belong to hyper-function and 150 samples belong to hypo-function of the thyroid gland dataset. The experiments were carried out using a Java application developed by the University of Waikato in New Zealand, Weka software. (http://www.cs.waikato.ac.nz/ml/weka). Multilayer perceptron and C4.5 are simple and efficient classifiers based on solid mathematical grounds. MLP has the remarkable ability train fast and extract patterns from complex data (Serpen, Jiang and Allred, 1997). In order to evaluate the efficiency of the method, classification accuracy has been utilized.

Results and discussion

This paper focuses on the comprehensive scalable analysis conducted on two datasets namely original and synthetic. In this, when feature selection is applied to the original dataset, the features were reduced from five to three (Jaganathan and Rajkumar, 2012). When applied to the synthetic dataset containing 25 features, the numbers of features selected are 13. These numbers of selected features have been used in the classification algorithms to obtain the accuracy in diagnosis. Table 1 describes the detailed accuracies obtained using ten-fold cross validation technique and various training-test partitions. It also shows the number of features selected for classification. From the table, we understand that 60%-40% training-test partition has produced 94.41% of classification accuracy for C4.5 classifier. But MLP has produced 94.04% of classification accuracy for 50%-50% training-test partitions. Similarly, when the accuracy is computed using ten-fold cross validations, it is found to be 94.97% for C4.5 and 94.88 for MLP. Amongst all the results, ten-fold cross validation has generated good result with C4.5 classifier than MLP using ten-fold cross validation as well as various training-test partitions. This result obtained from synthetic dataset is compared with the results obtained from original dataset by Jaganathan et al., (93.49% for C4.5 and 92.09% for MLP) the performance seems to proportionately better. However the result may likely to vary when real time data is obtained from thyroid patients. In spite of this, it can be concluded that the feature selection method applied in this research has produced significant results for both smaller dataset and larger dataset.

Conclusion and future work

The research in this paper has attempted to evaluate the performance of feature selection method applied to smaller and larger datasets. The results obtained are promising and the dimensionality reduction technique has proved proportionately equivalent characteristic results for both datasets. Hence it is observed that this method of feature selection works better even when applied to large volume of data. However, this method can further be tested with different types of data and various sizes of database with different number of samples and different numbers of features ranging from small size to large sizes in future.

References

Cao, Bin, Shen, Dou, Sun, Jian-Tao, et al. Feature selection in a kernel space. In International conference on machine learning (ICML); 2007 June 20-24; Oregon, USA.

Delen D, Walker G, Kadam A. Predicting breast cancer survivability: A comparison of three data mining methods. Artificial Intelligence in Medicine. 20005 34(2): 113-127.

Esin Dogantekin, Akif Dogantekin, Dervi Avci. An automatic diagnosis system based on thyroid gland: ADSTG. Expert Systems with Applications, 2010 37(9): 6368-6372.

Esin Dogantekin, Akif Dogantekin, Dervi Avci. An expert system based on Generalized Discriminant Analysis and Wavelet Support Vector Machine for diagnosis of thyroid disease. Expert Systems with Applications. 2011 38(1): 146-150.

Feyzullah Temurtas. A comparative study on thyroid disease diagnosis using neural networks. Expert Systems with Applications, 2009 36(1): 944-949.

Hoshi. An analysis of thyroid function diagnosis using Bayesian-type and SOM-type neural networks', Chemical and Pharmaceutical Bulletin, 2005 53: 1570–1574.

Jaganathan P, Rajkumar N. An expert system for optimizing thyroid disease diagnosis, Int. J. of Computational Science and Engineering. 2012 7( ): 232-238.

Juanying Xie and Chunxia Wang. Using support vector machines with a novel feature selection method for diagnosis of erythemato-squamous diseases. Expert Systems with Applications. 2011 38(5):5809-5815.

Keles A, Keles A. ESTDD: Expert system for thyroid diseases Diagnosis. Expert Systems with Applications. 2008 34(1): 242–246.

Ozyılmaz, L. and Yıldırım, T. Diagnosis of thyroid disease using artificial neural network methods. In Proceedings of ICONIP'02 nineth international conference on neural information processing 2002 Orchid Country Club, Singapore.

Pasi L. Similarity classifier applied to medical data sets, 2004, 10 sivua, Fuzziness in Finland'04 .In International conference on soft computing 2004 Helsinki, Finland & Gulf of Finland & Tallinn, Estonia.

Pasi L, Leppalampi T. Similarity classifier with generalized mean applied to medical data. Computers in Biology and Medicine, 2006 36(9): 1026-1040.

Polat K, Sahan S, Gunes S. A novel hybrid method based on Artificial immune recognition system (AIRS) with fuzzy weighted preprocessing for thyroid disease diagnosis. Expert Systems with Applications. 2007 32: 1141–1147.

Serpen G, Jiang H, Allred L. Performance analysis of probabilistic potential function neural network classifier. In Proceedings of artificial neural networks in engineering conference 1997 St. Louis, MO. 7: 471–476.

UCI Machine Learning Repository.[online] Available at :http://archive.ics.uci.edu/ml/

WEKA.[online] Available at: http://www.cs.waikato.ac.nz/ml/weka

Zhang G, Berardi L. V. An investigation of neural networks in thyroid function diagnosis. 1998 Health Care Management Science.29–37.

Table 1: Results obtained using synthetic dataset

Algorithm:	Synthetic Dataset			Accuracies obtained using 10-fold CV & various training-test partitions (%)				
	Instances	Features before selection	Features after selection	Ten-fold CV	50-50	60-40	70-30	80-20
C4.5	1075	25	13	94.97	94.22	9 4.41	93.78	93.95
MLP	1075	25	13	94.88	94.04	93.48	92.85	92.55