An International Journal
JSCDSS
http://www.jscdss.com
Vol.4 No.1 February 2017: 16-18
Article history:
Accepted 10 January 2017
Published online 11 March 2017

Journal of Soft Computing and Decision Support Systems

E-ISSN: 2289-8603

A Soft Computing Method for Mesothelioma Disease Classification

Mehrbakhsh Nilashi [a,*], Hossein Ahmadi [b], Leila Shahmoradi [b,*], Maryam Salahshour [a], Othman Ibrahim [a]
[a] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
[b] Health Information Management Department, 5th Floor, School of Allied Medical Sciences, Tehran University of Medical Sciences, No #17, Farredanesh Alley, Ghods St, Enghelab Ave, Tehran, Iran

*** Corresponding authors email addresses**: nilashidotnet@hotmail.com; lshahmoradi@tums.ac.ir

Abstract

Malignant Mesothelioma (MM) is a rare but highly aggressive tumour. The aim of this study is to improve the classification accuracy of MM disease by developing an intelligence system using machine learning techniques. Our method is developed through clustering, noise removal and classification approaches. Accordingly, we use Expectation Maximization (EM), Principal Component Analysis (PCA) and Support Vector Machine (SVM) for clustering, noise removal and classification tasks, respectively. We also develop the proposed method for incremental situation by applying the incremental PCA and incremental SVM for incremental learning of data. Experimental results on a malignant pleural mesothelioma disease dataset show that proposed method remarkably improves the accuracy of prediction and reduce computation time in relation to the non-incremental approaches. The hybrid intelligent system can assist medical practitioners in the healthcare practice as a decision support system.

Keywords: Malignant mesothelioma, Clustering, Incremental PCA, Incremental SVM, Machine Learning.

1. Introduction

Malignant Mesothelioma (MM) is an aggressive cancer of the serous membranes with a poor prognosis (Gemba et al., 2013; Spugnini et al., 2006). MM is a fatal tumor originating from the mesothelial tissue. There are two major localizations of MM: the pleura and peritoneum. The pleural form of malignant mesothelioma is the most common type accounting for more than 70% of all mesothelioma cases (Suzuki, 1981). Exposure to asbestos is a risk factor independent of tumor localization. The development of a mesothelioma is a lengthy process, the tumor appearing 25-60 years after asbestos exposure. There is no definitive standard of care for this disease and it has been shown that individual modalities such as chemotherapy, radiotherapy and surgery have been failed to prolong survival (Gemba et al., 2013).

This research aims to develop an intelligent system for Mesothelioma disease diagnosis using machine learning methods. Moreover, since in medical datasets constantly new information is available, hence, it is desirable to incrementally update the once trained models to reduce computation time in classifying the data. The proposed method in the study at hand supports incremental updates and re-learning of data and is more efficient in memory requirement. Overall, in comparison with research efforts found in the literature, in this research:

- Expectation Maximization (EM) is used for data clustering the data (Nilashi et al., 2015b; Nilashi et al., 2017; Nilashi et al., 2016a; Nilashi et al., 2016b; Nilashi et al., 2016c; Nilashi et al., 2016d).
- Support Vector Machine (SVM) is used for data classification (Nilashi et al., 2016d).
- Principal Component Analysis (PCA) is used for dimensionality reduction and dealing with the multi-collinearity problem in the experimental data (Nilashi et al., 2015a, Nilashi et al., 2015c).
- Incremental techniques, Incremental SVM (ISVM) and Incremental PCA (IPCA), are used for incremental learning.

Our study at hand is organized as follows: In Section 2, the research methodology is explained. In Section 3, the evaluation of method is presented. Finally, we conclude our work in Section 4.

2. Methodology of research

In the present study, PCA, EM, and SVM methods are used. In this study, EM clustering is used as an unsupervised classification method to cluster the data of experimental dataset into similar groups. We propose to rely on SVM to learn the classification models. One reason for this choice is that SVM and incremental SVM has shown higher predictive accuracy and lower computation

time in relation to other unsupervised machine learning techniques. We also use PCA for dimensionality reduction because the greatest source of difficulties in using classification methods is the existence of multi-collinearity in many sets of data. Furthermore, in this study, we apply PCA and SVM incrementally for newly arriving data to directly be incorporated into the models without retraining whole data. Therefore, we apply the incremental PCA and incremental SVM for online learning. In general, our methodology comprises two main phases which are:

Phase 1: In the first phase, the data is pre-processed in the first step (1). In the second step, EM clustering processing steps are performed to cluster the data (2) and then we apply PCA to reduce the dimensionality of the data and filter out potential noise (3). Next, classification models are learned by SVM for each cluster (4).

Phase 2: During this phase, i.e., after the system has been initially trained and deployed, on arrival of a new data sample, the models are incrementally updated. When a new data arrives, it is first added to the set of all data in the dataset. Next, we compute the distance of the new data point to the center of each cluster learned in the first phase. To do so, we use the Euclidian distance. Once the nearest cluster is determined, we apply the incremental PCA and incremental SVM methods to finally update the classification models of the corresponding cluster.

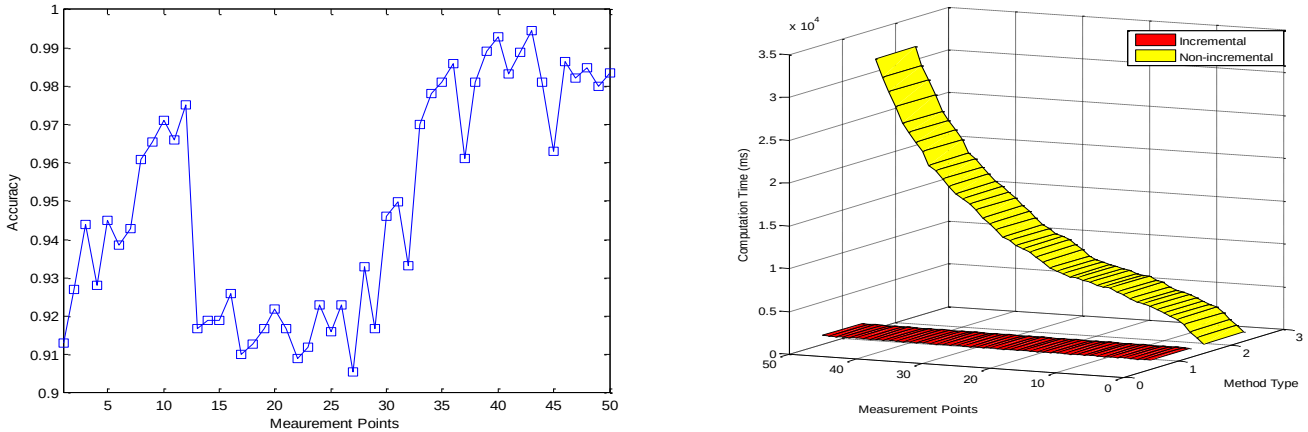
2.1 Dataset

In order to evaluate the proposed method, we have performed several experiments on a real-word dataset of Mesothelioma's disease. The dataset was retrieved from Data Mining Repository of the University of California, Irvine (UCI).

3. Results

To experimentally show the effectiveness of clustering and incremental approach (ISVM), we conduct the experiments on the public Mesothelioma dataset and compare with the methods of the non-incremental learning for computation time. It should be noted that the kernel parameters and penalty parameter C have been determined by 10-fold cross-validation.

In Fig. 1a, the classification accuracy of ISVM measured by ROC in each cluster for Mesothelioma is presented. From all plots in Fig. 1a, we can see the influence of using ISVM on accuracy is significant and the incremental update has provided a good classification accuracy measured by ROC in each cluster. The average accuracy obtained by the proposed method is about 94.95% for all clusters. It should be noted that the increment ratio for ISVM is considered 1.5% of incremental set and added to training set and we calculated accuracy in each fold of 10-fold cross validation. Fig. 1b presents the computation time results of our experiments for proposed method in the incremental situation. The computation time is plotted as a function of the incremental data percentage. As the figure show, non-incremental methods perform poor with respect to time for Mesothelioma dataset. From the curves as shown in the figures, it can be also observed that by increasing the number of incremental data the computation time is slightly raised. We compare the accuracy of our proposed method with the classification accuracy of the methods SVM, Linear Discriminant Analysis (LDA), Decision Tree (DT) and K-Nearest Neighbors (KNN) for Mesothelioma dataset. The performance of the classifiers that were compared with our method (IPCA-EM-ISVM) is shown in Table 1. From the results shown in this table, our proposed method proves to have a better accuracy for EM (0.9495) in relation to the other classification systems. Compared to SVM (87.53%), LDA (83.81%), DT (81.19%) and KNN (76.34%), our classification, clustering and noise removal techniques help to improve the classification accuracy of Mesothelioma's disease by more than 7.42%, 11.14%, 13.76% and 18.61%, respectively. This shows the effectiveness of incorporating the clustering and PCA techniques for the classification accuracy of Mesothelioma's disease.



(a) **(b)**

Fig. 1. Incremental SVM evaluation for (a) accuracy and (b) computation time

Table 1
Comparison of proposed method with other classifiers for Mesothelioma dataset

Method	Accuracy
KNN	76.34%
DT	81.19%
LDA	83.81%
SVM	87.53%
IPCA-EM-ISVM	94.95%

4. Conclusion and future work

In this paper, we propose a new hybrid intelligent system for Mesothelioma disease classification using machine learning techniques. We applied EM clustering algorithm to cluster the experimental Mesothelioma disease dataset and SVM for classification of disease types. Furthermore, PCA was used for dimensionality reduction and to address multi-collinearity in the dataset. Furthermore, since new information is constantly available in medical datasets, it is desirable to incrementally update the trained models to reduce the computation time. The proposed method in this study at hand then supports incremental updates that were more efficient in memory requirement. In order to analyse the effectiveness of the proposed method and validate the system, several experiments were conducted on Mesothelioma dataset. The dataset was taken from Data Mining Repository of the University of California, Irvine (UCI). The results indicated that the method which combines clustering, incremental PCA (IPCA) and incremental SVM (ISVM) obtain good classification accuracy and significantly reduce the computation time in relation to the non-incremental methods. All of the approaches used in this study, may also be applicable to other classification problems within the medical domain. However, there is still plenty of work in conducting researches on incremental algorithms for disease diagnosis in order to exploit all their potential and usefulness. In the future work, more attention should be paid to the datasets for disease classification and prediction using the incremental machine learning approaches. Hence, in our future study, we plan to evaluate the proposed method on additional datasets and in particular on large datasets to show the effectiveness of the incremental methods on computation time of large data in relation to the non-incremental ones.

References

Gemba, K., Fujimoto, N., Aoe, K., Kato, K., Takeshima, Y., Inai, K., & Kishimoto, T. (2013). Treatment and survival analyses of malignant mesothelioma in Japan. Acta Oncologica, 52(4), 803-808.

Nilashi, M. (2016a). An Overview of Data Mining Techniques in Recommender Systems. Journal of Soft Computing and Decision Support Systems, 3(6), 16-44.

Nilashi, M., bin Ibrahim, O., Ithnin, N., & Sarmin, N. H. (2015b). A multi-criteria collaborative filtering recommender system for the tourism domain using Expectation Maximization (EM) and PCA–ANFIS. Electronic Commerce Research and Applications, 14(6), 542-562.

Nilashi, M., Esfahani, M. D., Roudbaraki, M. Z., Ramayah, T., & Ibrahim, O. (2016c). A multi-criteria collaborative filtering recommender system using clustering and regression techniques. Journal of Soft Computing and Decision Support Systems, 3(5), 24-30.

Nilashi, M., Ibrahim, O. B., Ithnin, N., & Zakaria, R. (2015a). A multi-criteria recommendation system using dimensionality reduction and Neuro-Fuzzy techniques. Soft Computing, 19(11), 3173-3207.

Nilashi, M., Ibrahim, O. B., Mardani, A., Ahani, A., & Jusoh, A. (2016b). A soft computing approach for diabetes disease classification. Health Informatics Journal, 1460458216675500.

Nilashi, M., Ibrahim, O., & Ahani, A. (2016d). Accuracy Improvement for Predicting Parkinson's Disease Progression. Scientific Reports, 6.

Nilashi, M., Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). A knowledge-based system for breast cancer classification using fuzzy logic method. Telematics and Informatics, 34(4), 133-144.

Nilashi, M., Jannach, D., bin Ibrahim, O., & Ithnin, N. (2015c). Clustering-and regression-based multi-criteria collaborative filtering with incremental updates. Information Sciences, 293, 235-250.

Spugnini, E. P., Bosari, S., Citro, G., Lorenzon, I., Cognetti, F., & Baldi, A. (2006). Human malignant mesothelioma: molecular mechanisms of pathogenesis and progression. The international journal of biochemistry & cell biology, 38(12), 2000-2004.

Suzuki, Y. (1981). Pathology of human malignant mesothelioma. Semin Oncol, 8(3), 268-282.

JSCDS
E-ISSN: 2289-8603