Determination of Vital Cancer Sites in Malaysian Colorectal Cancer Dataset by Using A Fuzzy Feature Selection Method

Mohamad Faiz Dzulkalnine\textsuperscript{1}, Roselina Sallehuddin\textsuperscript{1}, Yusliza Yussof\textsuperscript{1}, Nor Haizan Mohd Radzi\textsuperscript{1}, Noorfa Haszlinna Binti Mustaffa\textsuperscript{1} and Lizawati Mi Yusuf\textsuperscript{1}

\textsuperscript{1}School of Computing, Universiti Teknologi Malaysia, Skudai, 81300 Johor, Malaysia.

Email: m.faiz.dzul@gmail.com

Abstract. In Malaysia, Colorectal Cancer (CRC) is one of the most common cancers that occur in both men and women. Early detection is very crucial and it can significantly increase the rate of survival for the patients and if left untreated can lead to death. With the lack of high-quality CRC data, expert systems and machine learning analysis are burdened with the presence of irrelevant features, outliers, and noise. This can reduce the classification accuracy for data analysis. Accordingly, it is essential to find a reliable feature selection method that can identify and remove any irrelevant feature while being resistant to noise and outliers. In this paper, Fuzzy Principal Component Analysis (FPCA) was tested for the classification of Malaysian’s CRC dataset. With the utilization of fuzzy membership in FPCA, the experimental results showed that the proposed method produces higher accuracy compared to PCA and SVM by almost 2\% and 5\% respectively. Empirical results showed that FPCA is a reliable feature selection method that can find the most informative features in the CRC dataset that could assist medical practitioners in making an informed decision.

1. Introduction

Colorectal cancer (CRC), also known as bowel cancer or colon cancer is among the most common cancer in the world. Currently, it is the third most common cancer and the fourth major cause of cancer deaths globally. Meanwhile, in Malaysia, CRC is the most prevalent type of cancer in men and only second in women behind breast cancer\textsuperscript{[1]}. According to the National Cancer Institute, the majority of the cases were of men and 66\% of them were diagnosed with the disease at a late stage (stage 3 or 4) \textsuperscript{[2]}. With the number of CRC incidence has been increasing up to four times since the past decades, this could be affecting Malaysia badly in the economic, healthcare, and social sector over time. One of the most effective ways to effectively reduce the mortality percentage is to have an early screening for colorectal cancer.

Therefore, it is important to determine the most important cancer sites in the colon that could influence the survival rate among colorectal cancer patients. Identification of these key cancer sites is vital as it can occur in various locations. Hence it could help medical practitioners making better and faster decisions making by prioritizing more on those key sites. However, with the limited amount of data, it is important to extract valuable information to be used in the physical process such as classification. Besides the lack of data, some of these data also contain irrelevant data. Studies showed when the number of features linearly increases, the required number of examples for learning increases exponentially and leads to the curse of dimensionality problems \textsuperscript{[3]}. Irrelevant features can reduce classification accuracy and increase the computational time of the SVM classifier \textsuperscript{[4]}. Therefore, the
main objective of this paper is to improve SVM classification performance by applying feature selection before the classification and identifying significant features in the CRC dataset.

A feature selection method is one of the well-known methods capable of removing irrelevant data in a dataset. Feature selection aims to improve the learning process, and prediction performance by selecting the best probable features, based on the characteristics of the dataset so that it can highlight the similarities and differences of the data. However, most feature selection methods such as Principal Component Analysis cannot handle outliers in a dataset and lead to inappropriate selections of significant features [5].

Outliers are extremely high or low values in the dataset which appears to be inconsistent with the rest of the data [6]. There are various causes of how outliers can present in a dataset such as human error or corruption in storage. This problem can also affect the accuracy performance of feature selection in classification as the features render irrelevant when outliers are present in the features.

The sensitivity to outliers can be diminished by incorporating fuzzy element in the calculation of the covariance matrix of the PCA. Improvement is due to the reason that feature space is divided differently as a result of nonlinearity in comparison to linear fuzzy PCA. By selecting the appropriate features with the consideration of outliers, there is no more issue with bias in classification. Thus, it is important to utilize a feature selection that can resist outliers to improve the classification accuracy of the classifiers.

As a result, a feature selection method nonlinear fuzzy robust principal component analysis based on fuzzy membership was proposed by Pasi Luukka in 2011 [7]. In this study, Luukka’s method is implemented in data pre-processing of the CRC dataset. In FPCA, the covariance matrix in FPCA is determined by using fuzzy clustering. This considers the presence of noise and outliers in the calculation of the mean. FPCA first determines the mean of the attributes by assigning them to a data cluster and noise cluster. The idea is to have a threshold that continuously influences the data by implements a noise cluster. The fuzziness variable, \( m \) determines the influence of data in the clusters. The higher the fuzziness variable, the sparser, and fuzzier the feature space of the clusters will become. The fuzziness variable is manually set by the user and Pasi Luukka suggested in the paper to set the fuzziness variable at 1.5 to achieve the best performance from FPCA.

Next, the distance of the data from the cluster center is determined to obtain a weight for the calculation of the mean. The weight exponent, \( w \) is another parameter that is given to different data to put more emphasis on their importance. The weight is randomly initialized and will keep updating until the objective function is minimized or until the iteration is maximum. In fuzzy membership, the closer the data to the cluster center with a high degree of membership, the higher the importance of the data, and will be given a much higher score. Outliers with a low degree of membership contribute to almost nothing thus lowering the PC score. The objective of this study is to demonstrate that by implementing fuzzy membership in a feature selection, a better set of features can be obtained and ultimately increase the classification accuracy by removing the issue of bias in classification. This paper is structured as follows. The next section is about the methodology of the study. Then, followed by the experiments and discussion about the results obtained. Finally, the conclusion is presented in section 4 which also includes future works.

2. Methodology

FPCA-SVM method consists of two phases, which is feature ranking by FPCA followed by classification by SVM. Figure 1 are an overall description of the phases followed by a more detailed description of the whole process.

Phase 1: Feature ranking by FPCA
i. Rank the features. First, the parameters for FPCA is set. The fuzziness variable, \( m \) is set at 1.5. In the first step, FPCA calculates the PC scores for each of the features. The relevancy of the feature is dependent on its PC scores. Higher PC scores signify higher relevancy while lower PC scores signify lower relevancy.
ii. Delete the lowest ranking feature. Before any features were deleted, the original dataset with all the features was classified by SVM so that the benchmark performance of the dataset is determined. Then, the lowest ranking feature is deleted. To determine whether the deleted feature is irrelevant, SVM will then classified the reduced dataset in the next steps.

**Figure 1.** Flowchart of FPCA-SVM.

**Phase 2: SVM classification**

i. SVM training. In SVM, there are two parameters \((c, g)\) that needed to be generated to ensure SVM performs optimally. Regularization parameter \(c\) determines the trade-off cost between minimizing the training error and the complexity of the model while \(g\), of the kernel function, defines the non-linear mapping from the input space to some high dimensional feature space. In this phase, 10-fold cross-validation is implemented to obtain the best parameters value. Cross-validation works by splitting the data into \(k\) subsets. Then, different combination value of the parameters is used to obtain cross-validation errors. The combination of \(c\) and \(g\) that produce the lowest cross-validation error will be selected for SVM training.

ii. SVM testing. After the training model is obtained, each of the testing set is then classified. Firstly, the feature with the lowest ranking is deleted. Then the remaining features will be classified by SVM. The accuracy obtained from the classification will be then compared with the benchmark performance. If the performance increase, this indicates that the deleted feature is irrelevant and does not contribute to the whole information of the dataset. This step repeats by deleting the next lowest ranking feature. The process stops when the classification accuracy either deteriorates or stagnant. The final classification accuracy is the previous value as the current accuracy is the result of the deletion of a relevant feature. The remaining feature subset is then considered as the most relevant and important features of the dataset.

3. Results and Discussions

In the next section, the dataset used in this study, and the performance measure used in this study will be discussed. Subsequently, the result that was obtained is also presented and discussed.

3.1 Experimental Data and Performance Measure

The medical dataset used in this study is obtained publicly from the website of Malaysia’s Department of Statistic (data.gov.my) in May 2019. This dataset contained information on the location for the colorectal cancer of the patients that were registered in the National Cancer Patient Registry - Colorectal Cancer [8]. This dataset has 12 features including the predictive feature with 1456 instances for each feature. The predictive feature of this dataset is the year the data recorded which is from 2008 to 2013 that is represented by “1” to “6” for each year. For example, 2008 represented by 1, 2009 by 2, and so on until 2013 which is represented by 6. All the experiments were implemented using Matlab software.
Table 1: Features in the colorectal dataset with the number of cancer occurrence per site.

| List of features (Number of reported cancer sites in patients from 2008-2013) |
|---------------------------------------------------------------|
| 1. Caecum (223), 2. Hepatic flexure (160), 3. Ascending colon (206) 4. Transverse colon (181), 5. Splenic flexure (103), 6. Descending colon (179), 7. Sigmoid colon (915), 8. Rectosigmoid (729), 9. Rectum (1456), 10. Anorectal (98), 11. Colon (98) |
| Total number of case (4348) |

To evaluate the performance of PCA and FPCA, SVM classification accuracy is used as the performance measurement method. The accuracy obtained from both of the feature selections is then compared. Equation (1) below is the formula for accuracy.

\[
\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)
\]

Where;
- True positives (TP) = Number of correct classification predicted as yes.
- True negatives (TN) = Number of correct classifications predicted as no.
- False-positive (FP) = Number of examples that are incorrectly predicted as yes when it is no.
- False-negative (FN) = No. of examples that are incorrectly predicted as no when it is actually yes.

The accuracy obtained from both of the feature selections is then compared. The higher accuracy indicates better performance and capabilities of the feature selection method.

3.2 Experimental Results

The results are tabulated in Table 2. In this study, FPCA is being compared with classical PCA and SVM. SVM is also used to establish the benchmark classification accuracy of the dataset without a feature selection method. In the table, the underlined and bold number signifies the deleted features from the final subset of the dataset.

Table 2. Classification accuracy for SVM, PCA, and FPCA for CRC dataset.

| Methods | Classification accuracy (%) | Feature selected | Number of features selected |
|---------|----------------------------|------------------|----------------------------|
| SVM     | 58.79                      | All              | 11                         |
| PCA     | 61.55                      | 1 > 10 > 5 > 7 > 4 > 8 > 11 > 3 > 6 > 9 > 2 | 10                         |
| FPCA    | 63.44                      | 9 > 8 > 11 > 5 > 10 > 6 > 7 > 2 > 4 > 3 > 1 | 9                          |

As shown in Table 2, the highest classification accuracy was obtained by FPCA with more than 63%. FPCA selects 9 number features from the whole dataset. Feature “Rectum” was ranked the highest signifies it is the most significant feature. PCA on the other hand rank the same feature in the 10th rank. This error in ranking affected SVM classification accuracy. PCA produces a slightly lower classification accuracy of 61.55% by selecting 10 features. The only feature PCA considered irrelevant and deleted was “Hepatic Flexure”. However, FPCA still considered the same feature relevant. It is also interesting that two features that were deleted by FPCA which are “Caecum” and “Ascending colon” were ranked highly by PCA. The former was the highest-ranking feature by PCA.

The difference in ranking is what increases FPCA classification accuracy as FPCA able to predict which feature contributes more to the overall information of the dataset. FPCA ranks differently...
compared to PCA because FPCA uses utilized a fuzziness variable, \( m \) determined based on the value of PC scores. The fuzziness variable will continuously affect the calculation of the PC scores, thus the ranking will change dramatically. As expected, the classification result attained from SVM using all the features were much worse compared to PCA-SVM and FPCA-SVM. As shown in Table 2, the classification accuracy of the original data was only 58.79%. SVM produces the expected results is due to the existence of irrelevant features which is proven to be able to affect the performance of the classifier.

4. Conclusions
In this paper, FPCA is compared with PCA to show that the fuzzy membership in FPCA able to select a better set of CRC cancer sites compared to classical PCA. The purpose of using FPCA instead of PCA is that its fuzzy membership makes the calculation more immune towards outliers. Therefore, FPCA can produce better learning and generalization ability in SVM classifier. Although the dataset used in this study did not have outliers, it can test the ranking ability of FPCA compared to PCA. Experimental results proved that the FPCA-SVM model yields the highest classification accuracy for the dataset tested. Thus proved that the prospect of fuzzy membership in the feature selection method and could be a more useful tool in improving the classification process in the future.

The proposed method will be a valuable tool to provide timely and robust analysis on the oncology practices such as safety and cost-effectiveness of treatment and most importantly the outcome of these patients. Future works of this study are to implement a fuzzy membership in other feature selection methods to further improve its immunity towards outliers and noise and to test the method on a dataset that has a high number of outliers.

References
[1] Loh S Y 2020 Journal of Oncology Research Review & Reports. 102 3
[2] Arunah Chandran D, Mustapha F I, and Hassan R A 2020 Med J Malaysia. 75(3) 235
[3] Marrir N, Wang H, Feng G, Li B, Jia M 2018 IEEE Access. 6 59657-59671
[4] Dzulkalnine M F, Saliehuddin R 2019 SN Applied Sciences. 1(4) 1-12
[5] Abbasi T, Lann K H, and Ismail I 2019 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS). 57-61
[6] Ottosen T B and Kumar P 2019 Environmental Science: Processes & Impacts. 21(4) 701-713.
[7] Luukka P 2011 International Journal of Fuzzy Systems. 13(3)
[8] Hassan M R, Khazim W K, Othman Z, Mustapha N R, Said R M, Leong T W 2014 The second report of the national cancer patient registry-colorectal cancer, 2008–2013. Kuala Lumpur, Malaysia

Acknowledgment
The authors wish to thank the Ministry of Higher Education, Malaysia, for funding this study under the Fundamental Research Grant Scheme FRGS/1/2019/ICT02/UTM/02/13. The authors would like to thank Research Management Centre (RMC) Universiti Teknologi Malaysia, UTM Zamalah Scholarship, and ALI@S for their support in research activities.