A hybrid approach to select features and classify diseases based on medical data

Hisham AbdelLatif¹,a, Jiawei Luo²,²

¹,²College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, China 410082.

a’heshamibb@yahoo.com, *Corresponding author: luojiawei@hnu.edu.cn

Abstract. Feature selection is popular problem in the classification of diseases in clinical medicine. Here, we developing a hybrid methodology to classify diseases, based on three medical datasets, Arrhythmia, Breast cancer, and Hepatitis datasets. This methodology called k-means ANOVA Support Vector Machine (K-ANOVA-SVM) uses K-means cluster with ANOVA statistical to preprocessing data and selection the significant features, and Support Vector Machines in the classification process. To compare and evaluate the performance, we choice three classification algorithms, decision tree Naïve Bayes, Support Vector Machines and applied the medical datasets direct to these algorithms. Our methodology was a much better classification accuracy is given of 98% in Arrhythmia datasets, 92% in Breast cancer datasets and 88% in Hepatitis datasets, Compare to use the medical data directly with decision tree Naïve Bayes, and Support Vector Machines. Also, the ROC curve and precision with (K-ANOVA-SVM) Achieved best results than other algorithms

1 Introduction

By using mining techniques and statistic tests, select the features only that have related to that diseases among all available features. By removing irrelevant attributes from the datasets, they found that improve the performance classification and the cost of classification also is reduced [9]. In this study, we aim to build a classifier for Medical data with the high accuracy using mining techniques and statistic tests. Many of the researchers were exposed to the problem of high-dimensional medical data, trying to reduce the high features, using many techniques [11]-[15]. This studies proposed some techniques such as GA, nearest neighbor method, Memetic Algorithms (MA), Ant Colony Optimization Algorithm (ACOFS), and others. To reduce the high dimensional in the datasets, we extract only the features that really related to the disease, based on the pre-processing for the datasets in our methodology. First part in this work shows an introduction to mining techniques and Medical informatics. Part two shows the related work. Part three shows the data description that used in this work. Part four illustrates the proposed methodology. Part five discusses the Results from our proposed methodology and compare our results to others classifiers, based on, Arrhythmia, Breast cancer and Hepatitis datasets. Part six, shows the conclusion associated with future work.
Related Work

Number of ways has been used to selecting informative Features to classify diseases, reduce data redundancy and improve classification process. In [1], proposed approach from combine both of filter and wrapper techniques to overcome the curse of high-dimensional, and extract only the desired features from datasets, also, aim to make the classification performance is more efficient and effective. In [2], implement the proposed method by three step, first, integrate between rules which describing the strings of medical data and decision tree algorithm to reduce the features. Then as in classification process, divided the dataset into into training and test sets. In the third step, the data set classified with the Artificial Immune Systems as the final steps in the approach. Prof.K.Rajeswari et al. [9], this approach is aim to support both of doctors and patients, by reduce the number of medical examinations, by extracting only the features that surely related to the diseases by using novel approach for extract the features and correlation to creates the association rules. Then features are used with some classifiers Neural Network Decision Tree SVM. Divya Chauhan et al. [10], the proposed classification model for the detection of Hepatitis diseases, based an efficient approach in which machine learning concepts are used.

Data Description

We tested our Method with three medical Datasets; these data were obtained from UCI Machine Learning Repository, as shown in Table1. Breast Cancer datasets comprise 357 patients diagnosed with breast cancer and 212 disease-free individuals with two classes. Arrhythmia datasets comprise 452 samples with 16 classes. Hepatitis comprises 155 samples with two classes. For more details about datasets in this link, http://archive.ics.uci.edu/ml/datasets.html.

| Features | Hepatitis | Arrhythmia | Breast Cancer |
|----------|-----------|------------|---------------|
| Attributes | 19 | 279 | 32 |
| instances | 155 | 452 | 569 |
| Class | 2 | 16 | 2 |

Proposed Classification Methodology

This proposed hybrid approach comprise three techniques, clustering, ANOVA and Support Vector Machine, with two main phases: the first phase is pre-processing datasets, second is the classification phase, as shown in Figure 1. In the first phase, a combine the k-means algorithm and ANOVA test are used to select the significant feature. In the second phase, SVM classifier is used to classify diseases based on the pre-processing of datasets

4.1 preprocessing phase

The aim of this phase to finding features that are significant and exclude others, this phase, consists of clustering by K-means algorithm and features selection by ANOVA test.
4.1.1 K-means

The concept of this algorithm, is easy both for interpretation and application, the main idea in k-means, is create a set of centroids, then assign each objects to the appropriate centroids, and so, the assign is continue to each cluster based on specific measures, even all objects are assign to the clusters. The centroids in the clusters must be smart way because of the different location for centroids generate the different result. So, put them as much as possible far away from each other is good way. Then, each point belonging to a given data set and assign it to the closest centroid. After, all points are assign to the cluster that means the cluster is completed and reached to the end.

Equation 1, shows the k-means algorithm and how this algorithm works.

\[
J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| \mathbf{x}_{i}^{(j)} - c_{j} \right\|^2
\]

(1)

The \( \mathbf{x}_{i}^{(j)} \) indicate to the points and \( c_{j} \) to the centers in cluster, and the minus between them in equation 1 is an indicate to difference of the points from their cluster centers. In this step, we have been set the number of clusters are three (\( k = 3 \)), for each data set. The k-means irregularly chooses the centroids for each cluster, then, calculate the Euclidean distance between each object and the centers, in all clusters based on equation 1. Based on the Euclidean distance for each objects, the k-means put the object in the cluster with the least distance, and so on until all objects assign to the three clusters.

4.1.2 ANOVA statistic

ANOVA test, t-test, Wilcoxon rank and others are methods applicable, can use them to assess the difference among the groups. Here, to obtain the only significant features from the clusters, we applied ANOVA test to each datasets, when (P-value) = 0.01. In nova test, must calculate the F-ratio as in equation 2, then comparison f-ratio with the p-value to select only the significant features and exclude others.

\[
F = \frac{(n - k) \sum n_{i} (\overline{Y}_{i} - \overline{Y}_{..})^2}{(k - 1) \sum (n_{i} - 1) s_{i}^2}
\]

(2)

Where \( \overline{Y}_{..} = \sum_{j=1}^{k} \frac{\overline{Y}_{ij} n_{i}}{n} \), \( \overline{Y}_{i} = \sum_{j=1}^{k} n_{i} \overline{Y}_{ij} / n \) and \( s_{i}^2 = \sum_{j=1}^{k} (Y_{ij} - \overline{Y}_{ij})^2 / (n_{i} - 1) \). \( \Sigma \) is indicate tithe sum is taken over the index [5].
4.2 building SVM classifier

This phase depend on the previous step, where we will takes the results from the pre-processing phase, and apply this results in develop SVM classifier and Classification of diseases, that we hope to make the classification process is more accuracy and computationally the fastest. Because the medical usually data suffer from high dimensional we use nonlinear Support Vector Machines, nonlinear uses a mapping to transform the original training data into a higher dimension. SVM has stand out as the technique of choice for modeling challenging high-dimensional data, compare to others techniques in the number of domain such as text mining [4], image mining [5], and bioinformatics [6].

We use Gaussian as kernel function with SVM Classifier to building developing our methodology. Mostly, with nonlinear problems, uses the Gaussian kernel function [7], because it like the sigmoid kernel for parameters also, it need few parameter than any others kernels. The x and y for the kernel function, that works as controls the complexity of the decision function versus the error minimization, can be defined by activating a two-dimensional grid search. For define the best pair of parameters, the performance for each combination must be calculated. The Tolerance in SVM use to stopping mechanism, tolerance indicates when the algorithm should be satisfied with the result and consider the building process complete.

We used the active learning with SVM, to generate a primary model on a small sample, number of the remaining training data, it is also working set the data record that is nearest to the decision boundary, and prune the decision hyperplane. Updates keep until the model reach on the training data or the maximum allowed number of support vectors.

The elect vectors in the pool are equal concerning their target values. With binary classification, half of the pool uses from each target class. When the data still is not enough for the target classes, the rest of the pool is used with other classes. While, in multi-class classification used a one-versus-all strategy.

5. Results and Discussion

Here, we discussed the results of the methodology K-ANOVA-SVM. Firstly, with k-means cluster, in each the process of clustering we set (number of clusters) K = 3 and maximum iteration = 100. Iteration process was uneven in each process of clustering but did not reach to 100. After clustering the datasets, we move to the second step was ANOVA test, we used it to features selection with p-value = 0.01.

From the result of k-means, we obtain three groups for each of these datasets. Arrhythmia dataset, G1=53 instances, G2 =170 instances and G3=229 instances, Hepatitis dataset G1=65 instances, G2 =37 instances and G3=48 instances and Breast cancer dataset G1=269 instances, G2 =150 instances and G3=150 instances. As we mentioned in the pre-processing step, we Applying ANOVA test to select only the significant features from clusters. With clusters from Arrhythmia dataset, we got 116 features as significant and, the rest are non-significant. Also from the clusters of breast cancer and Hepatitis, we got 6, and 8 significant features and the rest are non-significant from 32 and 19 features respectively as we can see in the table 2.

| Datasets    | Features | Significant features | P-Value |
|-------------|----------|----------------------|---------|
| Arrhythmia  | 267      | 116                  | 0.01    |
| Breast cancer | 32      | 6                    | 0.01    |
| Hepatitis   | 19       | 8                    | 0.01    |

Table 2: Feature Selection from datasets
The table 3, shows the proposed classifier is the better from others classifiers. Table 3, shows only the results with Arrhythmia datasets. As we see the highest accuracy we got when using (K-ANOVA-SVM), which is 98% with Arrhythmia datasets and 93%, 85% for the Breast cancer and Hepatitis respectively. The rest of the classifiers for the Arrhythmia datasets which we compared with are decision trees 64%, naïve Bayes is 83% with Breast cancer datasets, and SVM is73% with Arrhythmia datasets. Table 3, show the amount of the increase in the accuracy area under the curve (AUC) and precision in our method. Where the accuracy is 98%, AUC 91% and precision are 100

Table 3: Comparison DT, NB, SVM with K-ANOVA-SVM (Arrhythmia datasets)

| Method          | TPR | FPR | ACC | AUC  | Precision |
|-----------------|-----|-----|-----|------|-----------|
| DT              | 79  | 55  | 63  | 62   | 61        |
| NB              | 1   | 1   | 53  | 50   | 53        |
| SVM             | 61  | 79  | 70  | 75   | 78        |
| K-ANOVA-SVM     | 72  | 0   | 98  | 91   | 100       |

To evaluate the proposed classifier, we used the Receiver Operating Characteristics (ROC) analysis, ROC provides a means to compare individual models and determine thresholds which yield a high proportion of positive hits. Figure 2 illustrates ROC curve for Arrhythmia data classification also, shows the ROC curve measures the false and positive rate.

Here, the horizontal axis and vertical axis of a ROC measure the true positive rate (TPR) against the false positive rate (FPR). The top left-hand corner is the optimal location in a ROC curve, indicating best TPR, while the FPR is low. The area under the ROC measures the discriminating ability of the classification model. Figure 2: show ROC curve for the decision tree, Naive Bayes, Support Vector Machine (SVM) and (K-ANOVA-SVM) when dataset is Arrhythmia.

![ROC curve for DT, NB, SVM and K-ANOVA-SVM (Arrhythmia datasets)](image)

Fig. 2. ROC curve for DT, NB, SVM and K-ANOVA-SVM (Arrhythmia datasets)
In Figure 2 it is clear the K-ANOVA-SVM is better on each of the DT, NB, and SVM, which starts from zero and positive points draw a straight line to the top until 0.8 and then tends to the right draw negative points, as well as the area under the trend, is the best.

Table 4, shows the time of CPU time for K-ANOVA-SVM classifier versus SVM, NB and DT. From the table 4, it is clear building the K-ANOVA-SVM took the least time to build and test. While DT took the highest time, such as, in Arrhythmia data which was 75 seconds and 80 to build and test.

**Table 4: CPU time for build and test table (K-ANOVA-SVM) vs. SVM, NB and DT**

| Datasets      | CPU Time          |
|---------------|-------------------|
|               | K-ANOVA-SVM | SVM | DT | NB |
| Arrhythmia   | build 40s | test 45s | build 72s | test 75s | build 80s | test 68s | build 82s |
|             | 72s | 75s | 80s | 68s | 82s |
| Hepatitis    | build 11s | test 9s | build 19s | test 22s | build 22s | test 12s | build 12s |
| Breast Cancer| build 7s | test 12s | build 29s | test 10s | build 10s | test 12s | build 13s | test 18s |

6. Conclusions.
We developing a hybrid methodology to classify diseases, based medical datasets through combine three techniques, K-means, ANOVA test and Support Vector Machine. The Conclusion from this hybrid method explained from the compare results, how important and effective, the pre-processing for medical data, as well as the effective integration of both the K-means and ANOVA test, which had achieved the best results in classification process. With the K-ANOVA-SVM classifier, we obtained the highest accuracy, which was 98%, 93% and 85 with each datasets compared to three other classifiers: decision tree, Naïve Bayes, and Support Vector Machine algorithms. Also, CPU performance time was computed when using each algorithm for both buildings the classifier and testing the classifier.

Acknowledgements
This work is supported by Hunan Provincial Natural Science Foundation of China (Grant no.13JJ2017).

References
[1] Yonghong, P., Zhiqing W.U. and Jianmin J. A novel feature selection approach for biomedical data. journal of Biomedical Informatics, 2010(43) 15–23.
[2] K. Polat, S. Sahank, H. Kodaz, S. Günes. A New Classification Method for Breast Cancer Diagnosis: Feature Selection Artificial Immune Recognition System FS-AIRS [C]. International Conference on Natural Computation, 2005 (3611)830 – 838.
[3] Y. Saeys, I. Lanza, and P. Larrañaga. A review of feature selection techniques in bioinformatics [J]. bioinformatics, 2007 23(19) 2507–2517.
[4] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays [J]. PNAS, 1999 96(12) 6745–6750.
[5] Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J, Lai KM, Ji J, Dudoit S, Ng IO, Van De Rijn M, Botstein D, Brown PO. Gene expression patterns in human liver cancers [J]. Molecular
Biology of the Cell, 2002(13) 1929–1939.

[6] Hua J1, Xiong Z, Lowey J, Suh E, Dougherty ER. Optimal number of features as a function of sample size for various classification rules [J]. Bioinformatics, 2005(21) 1509–1515.

[7] Cristianini, N. and Shawe, T. J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods [B]. Cambridge University Press, 2000.

[8] Bharathi, A., M.Natarajan A. Cancer Classification of Bioinformatics data using ANOVA. International Journal of Computer Theory and Engineering, 2010 2(3).

[9] Prof.K.Rajeswari, Dr.V.Vaithiyanathan and Shailaja V.Pede. Feature Selection for Classification in Medical Data Mining. International Journal of Emerging Trends & Technology in Computer Science (IJJTCS), 2013,2(2) 492-497.

[10] Divya Chauhan and Varun Jaiswal. An efficient data mining classification approach for detecting lung cancer disease [C]. Communication and Electronics Systems (ICCES), International Conference on IEEE, 2017.1 - 8.

[11] Handels, H, Roß, Th, Kreusch, J, Wolff, H.H; Pöppl, S.J [J]. Feature selection for optimized skin tumor recognition using genetic algorithms. Artificial Intelligence in Medicine, 1999(16) 283-297.

[12] Younes Chtioui, Dominique Bertrand and Dominique Barba,"Feature Selection by a Genetic Algorithm.Application to Seed iscrimination by Artificial Vision," Journal of the Science of Food and Agriculture, 1998(76) 76-86.

[13] Yong Zhang, Dunwei Gong, Ying Hu, Wanqiu Zhang. Feature selection algorithm based on bare bones particle swarm optimization [J]. Neurocomputing, 2015 (148) 150-157.

[14] Cheng San Yang, Li Yeh Chuang, Yu Jung Chen, Cheng Hong Yang. Feature Selection using emetic Algorithms [C]. Third International Conference on Convergence and Hybrid Information Technology, 2008 416-423.

[15] Md. Monirul Kabir, Md. Shahjahan, Kazuyuki Murase. A new hybrid ant colony optimization algorithm for Expert Systems with Applications. Expert Systems with Applications, 2012(39) 3747-3763.