Comparación entre Support Vector Machine basado en Kmeans Clustering (SVM-Kmeans) y Support Vector Machine basado en Recursive Feature Elimination (SVM-RFE) para el diagnóstico de Cáncer de Mama

Comparison between Support Vector Machine based on Kmeans Clustering (SVM-Kmeans) and Support Vector Machine based on Recursive Feature Elimination (SVM-RFE) for Breast Cancer Diagnosis

Flor de Luz Palomino Valdivia[1], Herwin Alayn Huillcen Baca[2], Ivan Soria Solis[3]
[1]Professional School of Systems Engineering, José María Arguedas National University, Peru. fpalomino@unajma.edu.pe, 0000-0003-4638-5762
[2] Professional School of Systems Engineering, José María Arguedas National University, Peru. hhuillcen@unajma.edu.pe, 0000-0001-9385-7940
[3]Professional School of Systems Engineering, José María Arguedas National University, Peru. isoria@unajma.edu.pe, 0000-0003-4202-9402

Abstract
In this work, we emulate an architecture that combines the SVM classifier with the Kmeans clustering algorithm (SVM-KMEANS) and the method that uses SVM with Recursive Features Elimination (SVM-RFE), for breast cancer diagnosis. Both works were compared to assess the accuracy, taking two datasets for breast cancer: Breast Cancer Wisconsin Diagnostic dataset (WDBC), and Wisconsin Prognosis Breast Cancer (WPBC), obtained from UCI machine learning repository.
Taking into account only the WDBC dataset, it is observed that the method SVM-KMEANS, reaches the maximum level of accuracy (98.25%) with only 4 best features, while the method SVM-RFE reaches the same level of accuracy with 30 features. It is concluded that the SVM-KMEANS method is better than the SVM-RFE method.

Keywords: [Support Vector Machine (SVM)], [Recursive Feature Elimination(RFE)], [K-means Clustering] , [Chi-square] , [feature selection] , [Breast Cancer Diagnosis].

1. Introduction

In the medical area, the diagnosis of breast cancer is of great interest; since according to the World Cancer Report (US Cancer Statistics Working Group, 2019), it states that breast cancer is the second leading cause of death of women after lung cancer. Early and accurate diagnosis of breast cancer disease can lead to successful treatment.
There are many supervised learning classifiers (Lavanya & Rani, 2011), (Chunekar & Ambulgekar, 2009), that are introduced to help in disease diagnosis. It can be used Single classifiers, such as Naive Bayes (NB), Multilayer Perceptron (MLP), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) or Multiple classifiers that can be combined together to enhance the accuracy. In the last two decades, machine learning approaches have been well developed and applied in numerous areas.
Among many diagnosis and classification methods, SVM is a considerable method because has the ability to deal with very high dimensional data, and from computation perspective, provides a fast training process (Gürbüz & Kiliç, 2014). SVM provides good results in classification, but still needs more enhancement especially in disease diagnosis.
An effective method should not only tell the doctors the condition of a patient but also the most relevant features of one disease. Large amount of irrelevant and redundant data will not only affect the performance of classifier but also prolong the calculation time. Therefore, it is necessary to extract the most relevant features and employ them to construct the classifier; therefore, the proposal of both papers is the combination of the SVM classifier with other methods of feauture selection for breast cancer diagnosis.
In this paper, we emulate the architecture proposed by Gad, W. (Gad, 2016), which combines Support Vector Machine (SVM) classifier in conjunction with Kmeans clustering algorithm. In SVM-Kmeans, the clustering algorithm preserves the structure of the original dataset, and number of clusters is added to the training process. In addition, kernel and penalty factor parameters of SVM are defined as well. In clustering step, number of clusters k is usually defined by a domain expert. The proposed method aims at determining the number of clusters k. The unnecessary and irrelevant features are removed to speed up the computation time. Chi-square method, feature selection (Liu & Motoda, 2012) is adopted to select the most important features.
On the other hand, we emulate Support Vector Machine (SVM) classifier with recursive feature elimination (RFE), proposed by Yin, Z. (Yin et al., 2016). SVM-RFE uses the feature ranking to select the feature subset. It employs the discrimination function information of SVM to eliminate the feature with smallest correlation with the classifier from the original feature set. Repeating this operation until only one

feature remains, we obtain the ranking of the features. This work is based on the dataset WDBC downloaded from the repository of University of California at Irvine (UCI).
Both approaches were evaluated using two datasets for breast cancer: Breast Cancer Wisconsin Diagnostic dataset (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC) (Asuncion & Newman 2007), obtained from UCI machine learning repository.
We use Precision, Recall, and Accuracy performance to evaluate the results.
The proposed method by Gad, W. (Gad, 2016) is better than the method proposed by Yin, Z. (Yin et al., 2016).

2. Related Work

Wang, J. (Wang et al., 2005), they proposed speeds up the response of SVM classifiers by reducing the number of support vectors. This is done by the K-means SVM (KMSVM) algorithm. The KMSVM algorithm combines the K-means clustering technique with SVM using one more input parameter to be determined: the number of clusters. Experiments compare the KMSVM algorithm with SVM on real-world databases, and the results show that the KMSVM algorithm can speed up the response time of classifiers by both reducing support vectors and maintaining a similar testing accuracy to SVM.
In the paper of Santhanam (Santhanam & Padmavathi, 2015), K-Means is used for removing the noisy data and genetic algorithms for finding the optimal set of features with Support Vector Machine (SVM) as classifier for classification. The experimental result proves that, the proposed model has attained an average accuracy of 98.79 Lin, X., et. al. (2018)[12], they propose a method, SVM-RFE-OA, which combines the classification accuracy rate and the average overlapping ratio of the samples to determine the number of features to be selected from the feature rank of SVM-RFE. Meanwhile, to measure the feature weights more accurately, we propose a modified SVM-RFE-OA (M-SVM-RFE-OA) algorithm that temporally screens out the samples lying in a heavy overlapping area in each iteration. The experiments on the eight public biological datasets show that the discriminative ability of the feature subset could be measured more accurately by combining the classification accuracy rate with the average overlapping degree of the samples compared with using the classification accuracy rate alone, and shielding the samples in the overlapping area made the calculation of the feature weights more stable and accurate. The methods proposed in this study can also be used with other RFE techniques to define potential biomarkers from big biological data.

3. Methods

3.1. Support Vector Machine based on Kmeans Clustering (SVM-Kmeans)

We implement a program that replicates the method proposed by Gad, W. (Gad, 2016), SVM-KMEANS partitions data into k clusters, maintaining the main distributions of the dataset. Then, the important features are selected using Chi-square to reduce the large number of features. In the last step, SVM is applied. The proposed model consists of:

- Read breast cancer dataset.
- Preprocess dataset.
- Partition dataset into k clusters.
- Select important features using Chi-square method.
- Build SVM classifier.
- Select the best performance parameters.

3.1.1 Preprocessing

The data set must be in real number format. So in this preprocessing step categorical characteristics are transformed into numerical data. Then, the normalization function is performed (Graf et al., 2003).

$$F_{Normalization} = \frac{F - F_{min}}{F_{max} - F_{min}}$$

The breast cancer dataset used in classification experiments is the Breast Cancer Wisconsin Diagnostics(WDBC) and the Breast Cancer Wisconsin Prognosis(WPBC), which can be retrieved from UCI repository.

3.1.2 Clustering

K-means has aims at partitioning a given dataset into a certain number of clusters. It selects centroids randomly, and assigns data points to the nearest centroids to minimize the inter-cluster similarity.
Assume, there are n objects, Suppose that X and Z are two samples of pattern vectors, $O_1, O_2, O_3, \ldots, O_n$. Each object is a d-dimensional vector. Kmeans aims to partition the n objects into k sets $S = S_1, S_2, S_3, \ldots, S_n$ so as to minimize the within-cluster sum of squares, which is the sum of distance functions of each point in the cluster to the k center. Kmeans aims to minimize the objectives function, which is defined as:

$$argmin \sum_{a=1}^{k} \sum_{O \in S_i} \| O - \mu \|^2$$

where u_i is the mean of points in S_i.

3.1.3 Feature Selection

Chi-square is adopted to:
- Reduce training time.
- Reduce classification over fitting.
- Remove irrelevant features to solve dimensionality problem.

Chi-Square (X^2) is a statistical method to test independence between two features. Chi-Square is defined as:

$$X^2(t, c) = \sum_{e_t \in o,1} \sum_{e_c \in 0,1} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

where t is a feature in class c, N is the observed frequency, E the expected frequency. e_t equals 1 if the object contains a feature t and e_t equals 0 if the object does not contain t. e_c equals 1 if the object is in class c and e_t equals 0 if the object is not in class c.

3.1.4 SVM-Kmeans Classifier

SVM aims at maximizing the margin and the kernel trick to reach accuracy, and overcomes the problem curse of dimensionality. In

classification, SVM solves the quadratic optimization problem in equation:

$$min \parallel w \parallel^2 + C \sum_{i=1..l} \xi$$

The Gaussian radius basis function (RBF), polynomial function, and the sigmoid function are the most popular kernel functions. SVM is defined as

$$f(x) = sgn(\sum_{i=1..n} \propto_i K(x, x_i) + b)$$

where x_i is a support vector, α_i is the coefficient of the support vector, n is the number of support vectors, b is the bias, K is the kernel function, and sgn is the sign function. The proposed model adopts the Gaussian RBF kernel in equation:

$$K(x_1, x_2) = exp(-\gamma \parallel (x_1 - x_2) \parallel^2)$$

The penalty factor C, kernel parameter, ζ values are 1, 0.01, 1.0E-12 respectively. The test was carried out with the number of groups k at 2,4,5,6 and 8, where k=2 result having the best performance. Kmeans preserves the structure and distribution of the original data.

3.2. Support Vector Machine based with Recursive Feature Elimination aproach (SVM-RFE)

We also implement the method proposed by Yin, Z. (Yin et al., 2016). The main purpose of SVM-RFE is to compute the ranking weights for all features and sort the features according to weight vectors as the classification basis. SVM-RFE is an iteration process of the backward removal of features. Its steps for feature set selection are shown as follows.

- Use the current dataset to train the classifier.
- Compute the ranking weights for all features.
- Delete the feature with the smallest weight.

Implement the iteration process until there is only one feature remaining in the dataset;the implementation result provides a list of features in the order of weight. The algorithm will remove the feature with smallest ranking weight, while retaining the feature variables of significant impact. Finally, the feature variables will be listed in the descending order of explanatory difference degree. SVM-RFE's selection of feature sets can be mainly divided into three steps, namely, (1) the input of the datasets to be classified, (2) calculation of weight of each feature, and (3) the deletion of the feature of minimum weight to obtain the ranking of features [9].

4. Experiment ans Results

4.1. Used Dataset

Both approaches were evaluated using two datasets for breast cancer: Breast Cancer Wisconsin Diagnostic dataset (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC), obtained from UCI machine learning repository [8]. The WDBC dataset has 30 features and 569 records, while WPBC has 32 features and 198 records. The names of features is showed in Table 1:

Table 1: Feature names for WDBC and WPBC dataset

ID	WDBC	WPBC
1	radius_mean	radius_mean
2	texture_mean	texture_mean
3	perimeter_mean	perimeter_mean
4	area_mean	area_mean
5	smoothness_mean	smoothness_mean
6	compactness_mean	compactness_mean
7	concavity_mean	concavity_mean
8	concave points_mean	concave points_mean
9	symmetry_mean	symmetry_mean
10	fractal_dimension_mean	fractal_dimension_mean
11	radius_se	radius_se
12	texture_se	texture_se
13	perimeter_se	perimeter_se
14	area_se	area_se
15	smoothness_se	smoothness_se
16	compactness_se	compactness_se
17	concavity_se	concavity_se
18	concave points_se	concave points_se
19	symmetry_se	symmetry_se
20	fractal_dimension_se	fractal_dimension_se
21	radius_worst	radius_worst
22	texture_worst	texture_worst
23	perimeter_worst	perimeter_worst
24	area_worst	area_worst
25	smoothness_worst	smoothness_worst
26	compactness_worst	compactness_worst
27	concavity_worst	concavity_worst
28	concave points_worst	concave points_worst
29	symmetry_worst	symmetry_worst
30	fractal_dimension_worst	fractal_dimension_worst
31	-	tumor_size
32	-	lymph_node_status

Source: Own elaboration

4.2. Result for Support Vector Machine method

We test a program using Support Vector Machine for breast cancer classification without using a method for features reduction, the accuracy was 98.25% for WDBC dataset and 60% for WPBC dataset. The used parameters were:

- Kerbel : RBF
- Gamma : 0.0001
- C : 1
- Features : 30 for WDBC and 32 for WPBC dataset

4.3. Result for Support Vector Machine based on Kmeans Clustering (SVM-Kmeans)

The ranking of features for both datasets, using the chi-square method and adding two features, is shown in "Fig. 1" and "Fig. 2":

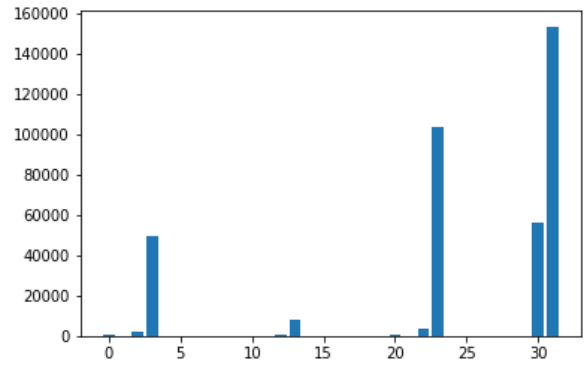

Fig. 1: Ranking of features for WDBC dataset
Source: Own elaboration

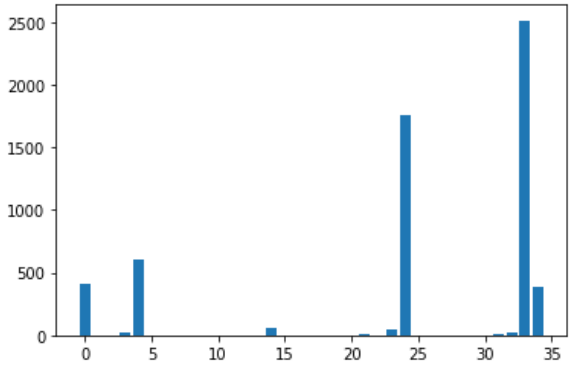

Fig. 2: Ranking of features for WPBC dataset
Source: Own elaboration

In both datasets, it is observed that the two added features have a good correlation with the class, likewise, the two best original features are features 3 and 23, corresponding to "area-mean" and "area-worst".
The results of accuracy is shown in the table 2:

Table 2: Results of accuracy with SVM-KMEANS method

	WDBC Dataset	WPBC Dataset
2 cluster and 2 best features	**96.49%**	**60.00%**
2 cluster and 4 best features	**98.25%**	**60.00%**
2 cluster and 8 best features	**98.25%**	**60.00%**
2 cluster and 16 best features	**98.25%**	**60.00%**
2 cluster and 30 best features	**98.25%**	**60.00%**
2 cluster and 32 best features	-	**60.00%**

Source: Own elaboration

4.4. Result for Support Vector Machine based with Recursive Feature Elimination aproach (SVM-RFE)

The results of accuracy is shown in the table 3:

Table 3: Results of accuracy with SVM-RFE method

	WDBC Dataset	WPBC Dataset
2 best features	**59.65%**	**60.00%**
4 best features	**59.65%**	**60.00%**
8 best features	**85.96%**	**60.00%**
16 best features	**92.98%**	**60.00%**
30 best features	**98.24%**	**60.00%**
32 best features	-	**60.00%**

Source: Own elaboration

5. Conclusions

The results obtained in both methods demonstrate that the WDBC dataset (569 samples) is more stable and consistent than the WPBC dataset (198 samples), due to the reduced number of samples. In our study we used 10% for tests; therefore working with the WPBC dataset (19 samples for testing), it does not necessarily reflect reliable accuracy results. It is shown that an accuracy of 60.00% is reached regardless of the number of reduced characteristics.
Taking into account only the WDBC dataset, it is observed that the method SVM-KMEANS, proposed by Gad, W. (Gad, 2016) reaches the maximum level of accuracy (98.25%) with only 4 best features, while the method proposed by by Yin, Z. (Yin et al., 2016) (SVM-RFE) reaches the same level of accuracy with 30 features. It is concluded that the SVM-KMEANS method is better than the SVM-RFE method.

Referencias

US Cancer Statistics Working Group, et al. United States cancer statistics: 1999–2013 cancer incidence and mortality data. Atlanta, GA: US *Department of Health and Human Services*, CDC, 2016.

Liu, H., & Motoda, H. (2012). Feature selection for knowledge discovery and data mining (Vol. 454). *Springer Science & Business Media*.

Asuncion, A., & Newman, D. (2007). UCI machine learning repository.

Chunekar, V. N., & Ambulgekar, H. P. (2009). Approach of neural network to diagnose breast cancer on three different data set. *ARTCom 2009 - International Conference on Advances in Recent Technologies in Communication and Computing*, 893–895. https://doi.org/10.1109/ARTCom.2009.225

Gad, W. (2016). SVM-Kmeans: Support Vector Machine based on Kmeans Clustering for Breast Cancer Diagnosis. *International Journal of Computer and Information Technology*, *05*(02), 2279–2764.

Graf, A. B. A., Smola, A. J., & Borer, S. (2003). Classification in a normalized feature space using support vector machines. *IEEE Transactions on Neural Networks*, *14*(3), 597–605. https://doi.org/10.1109/TNN.2003.811708

Gürbüz, E., & Kiliç, E. (2014). A new adaptive support vector machine for diagnosis of diseases. *Expert Systems*, *31*(5), 389–397. https://doi.org/10.1111/exsy.12051

Lavanya, D., & Rani, D. K. U. (2011). Analysis of Feature Selection with Classification : Breast Cancer Datasets. *Indian Journal of Computer Science and Engineering (IJCSE)*, *2*(5), 756–763. http://demo.pohonkeputusan.com/files/ANALYSIS OF FEATURE SELECTION WITH CLASSFICATION BREAST CANCER DATASETS.pdf

Santhanam, T., & Padmavathi, M. S. (2015). Application of K-Means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. *Procedia Computer*

Science, *47*(C), 76–83. https://doi.org/10.1016/j.procs.2015.03.185

Wang, J., Wu, X., & Zhang, C. (2005). Support vector machines based on K-means clustering for real-time business intelligence systems. *International Journal of Business Intelligence and Data Mining*, *1*(1), 54–64. https://doi.org/10.1504/IJBIDM.2005.007318

Yin, Z., Fei, Z., Yang, C., & Chen, A. (2016). A novel SVM-RFE based biomedical data processing approach: Basic and beyond. *IECON Proceedings (Industrial Electronics Conference)*, 7143–7148. https://doi.org/10.1109/IECON.2016.7793954