Data Analytics in Healthcare

Arun R
163190013

IE615: Data Analytics for Operations Research

November 23, 2017

Outline

Introduction

- The amount of data produced by healthcare industries grows exponentially.
- Bulk of the data comes from electronic health care, pharmacy, insurance claim, human tracking system and diagnostic instruments.
- The data can be leveraged using data analytics to provide better treatment to patients and reduce the operations cost.
- Healthcare data analytics can be used to,
  - Diagnose disease
  - Plan for disaster
  - Understand patient flow
  - Effectively manage resources and cost
  - Reduce fraud

Summary of Papers

Prediction	Feature Selection	ML Algorithm
Breast Cancer	F1-Score	SVM
Diabetes	Information Gain and SMOTE	Decision Trees, Logistic Regression, Naive Bayes and Random Forest
Hospital Readmission	Oversampling	Particle Swarm Optimization based SVM
Breast Cancer	K-Means	SVM

Problem Statement

- To diagnose breast cancer using machine learning techniques.
- Need a prediction model which is accurate and quick to build.
- Extract features from a given dataset using K-means clustering.
- Build a SVM-based prediction model on the extracted features.

Dataset: Instances

- Name: Wisconsin Diagnostic Breast Cancer (WDBC) dataset
- Date: November, 1965
- Number of instances: 569
- Number of instances in *benign tumor* class: 357
- Number of instances in *malignant tumor* class: 212

Dataset: Features

- Number of features: 30
- The features can be categorized as,
  - Radius
  - Texture
  - Perimeter
  - Area
  - Smoothness
  - Compactness
  - Concavity
  - Concave points
  - Symmetry
  - Fractal Dimension
- Mean, standard error and largest value are reported for each category.

Notation and Definition

K-means clustering is used to extract new features from the dataset. Notations used in this work are,

Notation	Definition
K	Number of clusters
F	Number of features in original dataset
N	Number of instances
S_c/S_k	Set of points in c^{th}/k^{th} cluster
X^i	i^{th} input in dataset
X_j^i	j^{th} feature in i^{th} input
X^{μ_k}	Center of k^{th} cluster
$X_j^{\mu_k}$	j^{th} feature of center of k^{th} cluster

Feature Extraction

- K-means clustering is used to find hidden patterns in each class.
- Cluster centers are used to extract new features.
- Validity ratio is used to fix the number of clusters in each class.

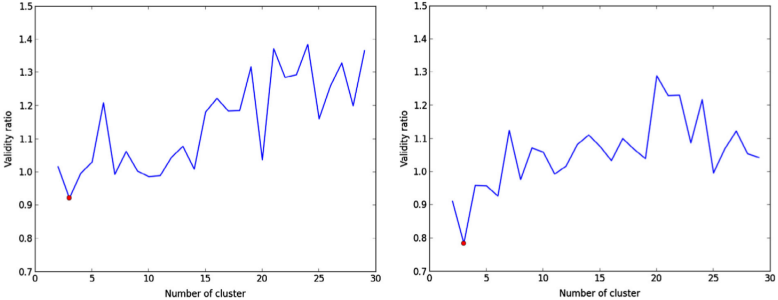
$$\text{Validity Ratio} = \frac{d_{avg}}{d_{min}}$$

where,

$$d_{avg} = \frac{\sum_{k=1}^{K} \sum_{i \in S_k} \sqrt{\sum_{j=1}^{F} \left(X_j^i - X_j^{\mu_k}\right)}}{N}$$

$$d_{min} = \min \left[\sum_{j=1}^{F} \sqrt{\left(X_j^{\mu_{k_2}} - X_j^{\mu_{k_2}}\right)^2} \right] \forall k_1 \neq k_2$$

Original Results



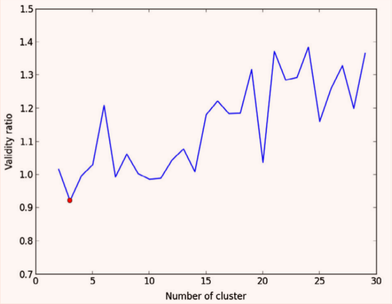
(a) Benign tumors (b) Malignant tumors

Figure: Variation of validity ratio with number of clusters

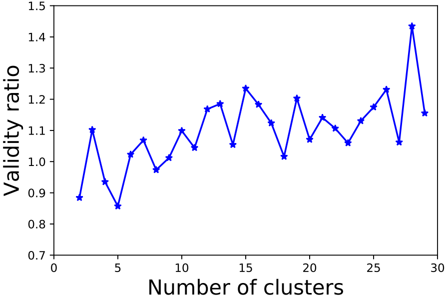
Optimal number of clusters is three for both the classes.

New Results (Benign): Full Normalization

Instances of both the classes are normalized together.



(a) Original Result

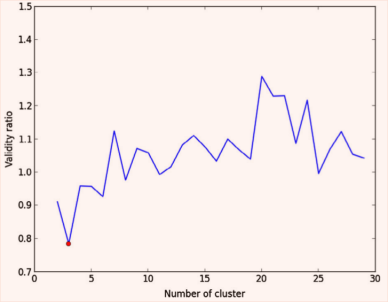


(b) New Result

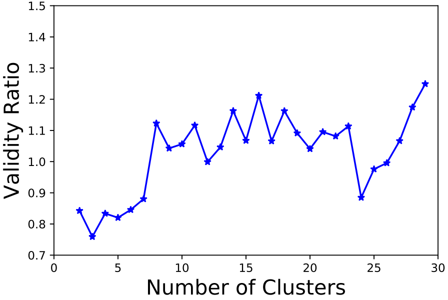
- Minimum value is achieved when $K = 5$. Results are not matching.
- Validity ratio is not same in both the results. Could be due to random initialization of cluster center in K-means.

New Results (Malignant): Full Normalization

Instances of both the classes are normalized together.



(a) Original Result

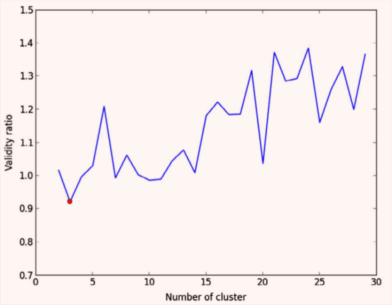


(b) New Result

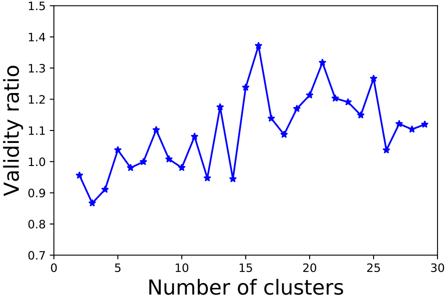
- Minimum value is achieved when $K = 3$.
- Validity ratio is not same in both the results. Could be due to random initialization of cluster center in K-means.

New Results (Benign): Separate Normalization

Instances of both the classes are normalized separately.



(a) Original Result

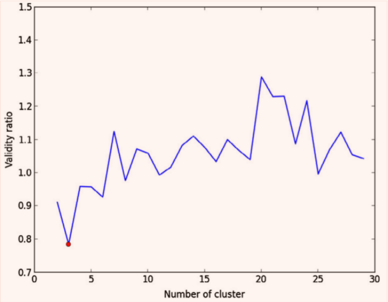


(b) New Result

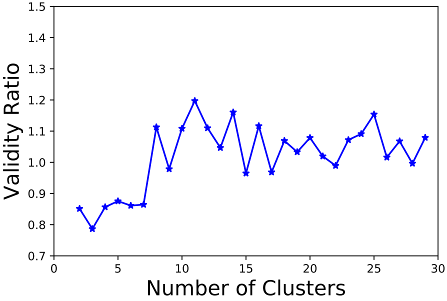
- Minimum value is achieved when K = 3.
- Validity ratio is not same in both the results. Could be due to random initialization of cluster center in K-means.

New Results (Malignant): Separate Normalization

Instances of both the classes are normalized separately.



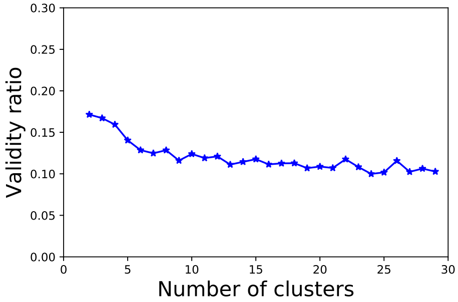
(a) Original Result



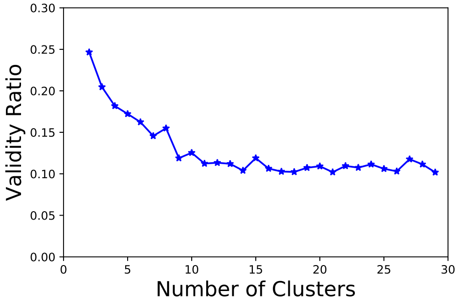
(b) New Result

- Minimum value is achieved when K = 3.
- All the other results are obtained from separate normalization.

New Results: Silhouette Value



(a) Benign Tumor



(b) Malignant Tumor

- For both the classes maximum value is achieved when $K = 2$.
- Different from results obtained using validity ration

Feature Extraction and SVM model

- The six cluster centers give symbolic representation of the clusters.
- Six features are extracted using these six cluster centers.

$$f_c(X_j^i) = \begin{cases} 1 - \frac{|X_j^{\mu_c} - X_j^i|}{\max |X_j^{\mu_c} - X_j^n|}, & \text{if } \min(X_j^n) \leq X_j^i \leq \max(X_j^n),\ \forall n \in S_c \\ 0, & \text{otherwise} \end{cases}$$

$$p_c = \frac{1}{F} \sum_{j=1}^{F} f_c(X_j^i), \ \ 1 \leq c \leq K^m + K^b$$

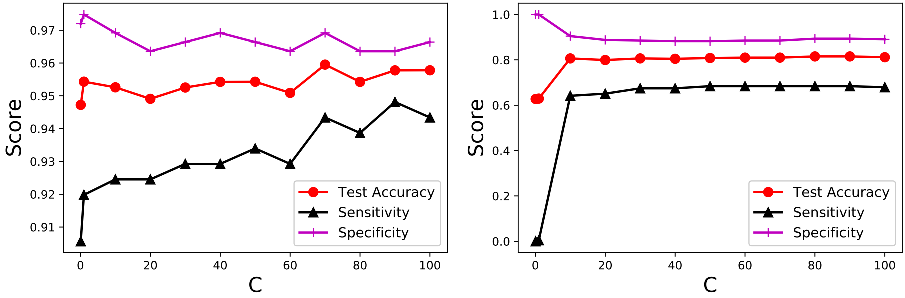
SVM model is built using the extracted features to diagnose breast cancer.

Experimental Setup

Parameter	Value
SVM penalty (C)	0.1, 1, 10, 20, $\cdots$, 100
Kernels	Linear and Sigmoid
Cross Vaidation	10-fold cross validation
Performance metrics	Test accuracy, sensitivity, specificity and time
Programming Language	Python (Scikit)
Processor	Intel Core i7 with 2.5 GHz processor

Linear Kernel: Accuracy, Sensitivity and Specificity

Y axis range is different in the figures



(a) SVM

(b) KSVM

- Highest accuracy for SVM is 0.96 at $C = 70$.
- Highest accuracy for KSVM is 0.81 at $C = 80$.

Linear Kernel: Confusion Matrix

Actual \ Prediction	Benign	Malignant
Benign	346	11
Malignant	12	200

(a) SVM ($C = 70$)

Actual \ Prediction	Benign	Malignant
Benign	315	42
Malignant	67	145

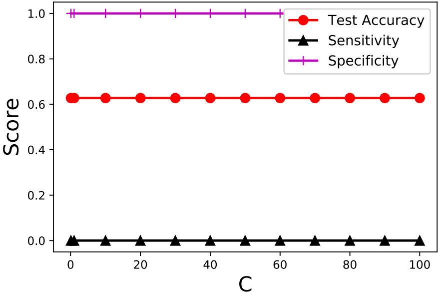
(b) KSVM ($C = 80$)

Linear Kernel: Computation Time

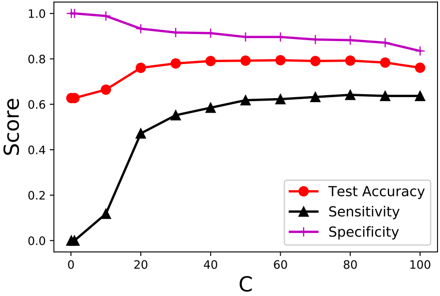
C	SVM Time (in sec)	KSVM Time (in sec)	C	SVM Time (in sec)	KSVM Time (in sec)
0.1	2.8	0.03	50	94.5	0.05
1	11.3	0.04	60	67.8	0.06
10	47.9	0.04	70	71.8	0.05
20	49.1	0.05	80	71.34	0.06
30	64.3	0.04	90	74.9	0.06
40	79.2	0.05	100	75.91	0.07

As expected, less computation time is required in KSVM than SVM.

Sigmoid Kernel: Accuracy, Sensitivity and Specificity



(a) SVM



(b) KSVM

- Highest accuracy for SVM is 0.63 at all C and gamma value.
- Highest accuracy for KSVM is 0.79 at C = 80 and gamma = 0.167.
- Accuracy for KSVM reported in paper is 0.97 (C value and gamma value are not mentioned).

Sigmoid Kernel: Confusion Matrix

Actual \ Prediction	Benign	Malignant
Benign	357	0
Malignant	212	0

(a) SVM (All C)

Actual \ Prediction	Benign	Malignant
Benign	320	37
Malignant	81	131

(b) KSVM ($C = 80$)

Paper Summary

Prediction	Feature Selection	ML Algorithm
Breast Cancer	F1-Score	SVM
Diabetes	Information Gain and SMOTE	Decision Trees, Logistic Regression, Naive Bayes and Random Forest
Hospital Readmission	Oversampling	Particle Swarm Optimization based SVM
Breast Cancer	K-Means	SVM

References I

Mehmet Fatih Akay. "Support vector machines combined with feature selection for breast cancer diagnosis". In: Expert systems with applications 36.2 (2009), pp. 3240–3247.

Manal Alghamdi et al. "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project". In: PLoS One 12.7 (2017), e0179805.

Bichen Zheng, Sang Won Yoon, and Sarah S Lam. "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms". In: Expert Systems with Applications 41.4 (2014), pp. 1476–1482.

Bichen Zheng et al. "Predictive modeling of hospital readmissions using metaheuristics and data mining". In: Expert Systems with Applications 42.20 (2015), pp. 7110–7120.