**MACHINE LEARNING PROJECT**

**Milestone #1**

**A. Dataset Building and Normalization**

1. **Source of Dataset**

* UCI Machine Learning Repository – Center for Machine Learning and Intelligent Systems http://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set
* Dr. Balint Antal, Department of Computer Graphics and Image Processing Faculty of Informatics, University of Debrecen, 4010, Debrecen, POB 12, Hungary antal.balint '@' inf.unideb.hu
* Dr. Andras Hajdu, Department of Computer Graphics and Image Processing Faculty of Informatics, University of Debrecen, 4010, Debrecen, POB 12, Hungary hajdu.andras '@' inf.unideb.hu

1. **Description of Dataset**

Number of original features: 20

Sample size: 1151

This dataset contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not. All features represent either a detected lesion, a descriptive feature of a anatomical part or an image-level descriptor. The underlying method image analysis and feature extraction as well as our classification technique is described in Balint Antal, Andras Hajdu: An ensemble-based system for automatic screening of diabetic retinopathy, Knowledge-Based Systems 60 (April 2014), 20-27.

1. **Name, Description and type of each feature**

A Result of Quality Assessment-----------------------------------------------------------------------------------Real = [0 - 1]

B Result of Pre-screening------------------------------------------------------------------------------------------ Real = [0 - 1]

C-H Results of MA detection----------------------------------------------------------------------------------------- Real = [0 - 1]

I-P Results of MA detection for exudates------------------------------------------------------------------------ Real = [0 - 1]

Q Euclidean distance of the center of the macula and the center of the optic disc----------------- Real = [0 - 1]

R The diameter of the optic disc--------------------------------------------------------------------------------- Real = [0 - 1]

S The binary result of the AM/FM-based classification----------------------------------------------------- Real = [0 - 1]

T Class label--------------------------------------------------------------------------------------------Binomial ⊆ [false, true]

The attributes were set from Numerical to Binomial, from Nominal to Numerical, and then normalized to attain a linear and more robust relationship.

1. **Processed/Normalized dataset**

https://www.dropbox.com/s/ukzgzmqoz1hm9ay/Milestone-1-diabetic-retinopathy-normalized.xlsx?dl=0

1. **Number of features that were removed and the reason for their removal**

None

**B. Performance of kNN, Decision Trees, Bayesian Network on Different Feature Sets**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | 61.25 % | 68.20 % | 63.34 % | 67.16 % |
| **kNN (k=3)** | 62.29 % | 71.42 % | 64.47 % | 64.64 % |
| **kNN (k=5)** | 63.16 % | 72.20 % | 66.20 % | 65.68 % |
| **kNN (k=7)** | 64.73 % | 70.89 % | 68.03 % | 67.33 % |
| **kNN (k=9)** | 63.94 % | 70.89 % | 65.94 % | 68.20 % |
| **Decision Trees** | 53.69 % | 54.39 % | 54.39 % | 53.26 % |
| **Bayesian Network (e.g. Naïve Bayes)** | 60.30 % | 65.68 % | 65.07 % | 59.77 % |

**Table 1 – Accuracy**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | 0.225 | 0.365 | 0.267 | 0.341 |
| **kNN (k=3)** | 0.247 | 0.429 | 0.291 | 0.292 |
| **kNN (k=5)** | 0.265 | 0.446 | 0.326 | 0.315 |
| **kNN (k=7)** | 0.296 | 0.420 | 0.362 | 0.348 |
| **kNN (k=9)** | 0.281 | 0.422 | 0.321 | 0.366 |
| **Decision Trees** | 0.069 | 0.030 | 0.093 | 0.017 |
| **Bayesian Network (e.g. Naïve Bayes)** | 0.233 | 0.322 | 0.313 | 0.201 |

**Table 2 – Kappa**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | 0.500 | 0.500 | 0.500 | 0.300 |
| **kNN (k=3)** | 0.660 | 0.736 | 0.675 | 0.684 |
| **kNN (k=5)** | 0.680 | 0.771 | 0.703 | 0.713 |
| **kNN (k=7)** | 0.694 | 0.761 | 0.717 | 0.725 |
| **kNN (k=9)** | 0.696 | 0.767 | 0.708 | 0.731 |
| **Decision Trees** | 0.549 | 0.548 | 0.565 | 0.529 |
| **Bayesian Network (e.g. Naïve Bayes)** | 0.689 | 0.678 | 0.700 | 0.652 |

**Table 3 – AUC**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | 64.40 % | 71.60 % | 66.67 % | 69.19 % |
| **kNN (k=3)** | 65.83 % | 75.09 % | 68.23 % | 67.59 % |
| **kNN (k=5)** | 66.91 % | 76.89 % | 70.48 % | 69.22 % |
| **kNN (k=7)** | 68.60 % | 75.75 % | 71.97 % | 71.10 % |
| **kNN (k=9)** | 67.88 % | 76.85 % | 70.02 % | 72.15 % |
| **Decision Trees** | 56.27 % | 53.80 % | 58.24 % | 53.54 % |
| **Bayesian Network (e.g. Naïve Bayes)** | 81.56 % | 73.28 % | 74.59 % | 64.45 % |

**Table 4 – Precision**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | 60.39 % | 66.45 % | 61.87 % | 68.74 % |
| **kNN (k=3)** | 60.23 % | 69.07 % | 61.87 % | 64.16 % |
| **kNN (k=5)** | 60.56 % | 68.09 % | 62.52 % | 63.67 % |
| **kNN (k=7)** | 61.87 % | 66.45 % | 65.14 % | 64.81 % |
| **kNN (k=9)** | 60.88 % | 64.65 % | 62.68 % | 65.20 % |
| **Decision Trees** | 57.28 % | 99.67 % | 49.75 % | 90.34 % |
| **Bayesian Network (e.g. Naïve Bayes)** | 32.57 % | 55.65 % | 51.88 % | 54.01 % |

**Table 5 – Recall**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Complete Feature Set** | **Feature Selection using Forward Search** | **Feature Selection using Backward Search** | **Top 9 features (with higher coefficients) based on 3 Components of Principal Components Analysis (PCA)** |
| **kNN (k=1)** | 62.33 % | 68.93 % | 64.18 % | 68.97 % |
| **kNN (k=3)** | 62.91 % | 71.95 % | 64.89 % | 65.83 % |
| **kNN (k=5)** | 63.57 % | 72.22 % | 66.26 % | 66.33 % |
| **kNN (k=7)** | 65.06 % | 70.79 % | 68.38 % | 67.81 % |
| **kNN (k=9)** | 64.19 % | 70.22 % | 66.15 % | 68.56 % |
| **Decision Trees** | 56.77 % | 69.88 % | 53.66 % | 67.23 % |
| **Bayesian Network (e.g. Naïve Bayes)** | 46.55 % | 63.26 % | 61.20 % | 58.77 % |

**Table 6 – F-Measure**

**C. Discussion of Results**

Based on the results, the following characteristics were observed:

1. In the Complete Feature Set, kNN (k=7) was the best classifier.
2. In the Feature Selection Forward Search, kNN (k=5) was the best classifier.
3. In the Feature Selection Backward Search, kNN (k=7) was the best classifier.
4. In PCA, kNN (k=9) was the best classifier.

This shows that overall, kNN (k=7) was the best classifier.

**D. Problems Encountered**

None