DEVELOPMENT OF A PARTIAL SUPERVISION STRATEGY TO AUGMENT A NEAREST NEIGHBOUR CLUSTERING ALGORITHM FOR BIOMEDICAL DATA CLASSIFICATION

Sameh A. Salem, Nancy M. Salem and Asoke K. Nandi

Signal Processing and Communications Group
Department of Electrical Engineering and Electronics, The University of Liverpool
Brownlow Hill, L69 3GJ, Liverpool, UK

Keywords: Data clustering, Partial supervision, Retinal blood vessels segmentation, Breast cancer classification.

Abstract: In this paper, a partial supervision strategy for a recently developed clustering algorithm NNCA (Salem et al., 2006), Nearest Neighbour Clustering Algorithm, is proposed. The proposed method (NNCA-PS) offers classification capability with smaller amount of a priori knowledge, where a small number of data objects from the entire dataset are used as labelled objects to guide the clustering process towards a better search space. Results from the proposed supervision method indicate its robustness in classification compared with other classifiers.

1 INTRODUCTION

Data clustering is a common technique for data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Clustering is the grouping of individuals in a population in order to discover structures in the data. In some sense, we would like the individuals within a group to be close or similar to one another, but dissimilar from the individuals in the other groups (Webb, 2003; Theodoridis and Koutroubhas, 2003). Recently, a number of clustering algorithms has been proposed. The basic two types of clustering algorithms are partitional and hierarchical algorithms. Their main purpose (Xu and Wunsch, 2005; Jain et al., 1999; Jain and Dubes, 1988) is to evolve a $N_C \times n$ partition matrix $U(X)$ of a dataset $X \{x_1, x_2, \ldots, x_n\}$ in $R^n$, representing its partitioning into a number of $N_C$ clusters $(C_1, C_2, \ldots, C_{N_C})$. The partition matrix $U(X)$ may be represented as $U = [u_{m,j}], m = 1, \ldots, N_C$ and $j = 1, \ldots, n$, where $u_{m,j}$ is the membership of pattern $x_j$ to cluster $C_m$. In hard partitioning of the data, the following conditions hold: $u_{m,j} = 1$ if $x_j \in C_m$, otherwise, $u_{m,j} = 0$.

Clustering is unsupervised classification where there are no predefined classes (labels) and no a priori knowledge of the data, while supervised classification requires a complete knowledge of the data where the class label and the number of classes (labels) are predefined (Bouchachia and Pedrycz, 2006). The process of labeling data objects is always an expensive and error-prone task that requires time and human intervention. In many situations, objects are neither perfectly labelled nor completely labelled. Therefore, the main idea of clustering with partial supervision strategy is to take the advantage of the smaller proportion of labelled objects to guide the clustering process of the unlabelled objects.

One of the typical applications of clustering with partial supervision is Computer-Aided Diagnosis (CAD) which has become one of the major research subjects in medical imaging and diagnostic radiology (Doi, 2005). The basic concept of CAD is to provide a computer output as a second opinion to assist radiologists’ image interpretation by improving the accuracy and consistency of radiological diagnosis (Doi, 2005). The design of clustering with partial supervision in CAD can play an important role in improving CAD performance at small amount of knowledge, where only some labelled objects or regions of an image can assist in identification of any suspicious objects or regions.

This paper proposes a novel partial supervision strategy for the recently developed clustering algorithm NNCA (Salem et al., 2006). We examine its applicability and reliability using datasets from real-world problems, where the proposed method is used to segment the blood vessels in retinal images which
can help in early detection and diagnosis of many eye diseases, and it is used to classify breast tumors into either malignant or benign. Additionally, this paper presents a comparative evaluation of the proposed algorithm with some other algorithms.

## 2 THE NNCA CLUSTERING ALGORITHM

**NNCA** (Salem et al., 2006) is a modified version of the **KNN** classifier, and it is divided into two stages for creating **NC** clusters. First stage is to select **N** objects randomly. Then non-overlapping clusters are created from these **N** objects, each of maximum size **Kinit** objects (the choice of **Kinit** ensures that more than **NC** clusters are generated here). Afterwards an iterative control strategy is applied to update the clusters and their memberships by increasing the number of neighbours until the number of non-overlapping clusters are created. Second stage is to cluster the remaining objects. For each unclustered object **q**, **K** nearest clustered objects are found. Then, the cluster to which most of these **K** clustered objects belong is deemed to be one to which the object **q** belongs to.

The **NNCA** clustering algorithm is detailed in Algorithm 1. Let each object **x** be described by the feature vector:

\[
< a_1(x), a_2(x), \ldots, a_p(x) >
\]

where **a_r(x)** is used to denote the values of the **r**-th attribute of data point **x**. If we consider two objects **x_i** and **x_j**, then the distance between them is defined as

\[
d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (a_r(x_i) - a_r(x_j))^2}
\]  

(1)

A fuzzy clustering, where all objects are allowed to belong to all clusters with different degrees of membership, is achieved by obtaining the mean value of the **K** nearest neighbours for each object in the dataset. Therefore, hard partition as well as soft partition can be obtained. For an object **x_q** to be clustered, let **x_1 \ldots x_N** denote the nearest **K** clustered objects to **x_q** and **C(x_q) \in \{1, \ldots, NC\}** is the cluster index for object **x_q**. Hard partition value for **x_q** is:

\[
C(x_q) = \arg \max_{n=NC} \sum_{r=1}^{K} \delta(n - C(x_r)),
\]  

(2)

and soft partition vector is:

\[
C(x_q) = \frac{\sum_{r=1}^{K} \delta(C(x_r) - C(x_q))}{K}
\]  

(3)

**Algorithm 1** Nearest Neighbour Clustering Algorithm (Salem et al., 2006)

**Input** (data, **N**, **Kinit**, **NC**, **K**) where:

* **N** is the number of random objects to be clustered.
* **Kinit** is the nearest neighbour objects from **N**.
* **NC** is the user defined number of clusters.
* **K** is the number of nearest clustered objects.

**# Step 1:** Create **Nc** non-overlapped clusters

**# (a) Create initial clusters:**

* Initially, all the **N** objects are unclustered.
  let **M** = 1
  **For** **i** = 1 to **N**
  **IF** (object **i** is unclustered)
  - Assign **i** and its unclustered neighbours (from **N**) of the **Kinit** nearest neighbours to cluster # **M**.
  - **M** = **M** + 1
  **End IF**
  **End For**

**# (b) Merge clusters:**

* **DO**
  - **Kinit** = **Kinit** + 1
  - Assign each clustered object to the common cluster of the **Kinit** nearest neighbours.
  - Update the number of clusters \( \rightarrow M \)
  **WHILE** (**M** > **NC**)  

**# Step 2:** Find the nearest **K** neighbours for each remaining object

* Assign each unclustered object to the common cluster of the **K** nearest clustered objects.
* Use Eq. 2 to find hard partition and Eq. 3 to find soft partition.

**Output** (Hard partition vector, Soft partition matrix)

Figure 1 shows a sub-image from a colour retinal image and its ground truth along with the corresponding segmented sub-images after applying **NNCA**.

## 3 NNCA WITH PARTIAL SUPERVISION STRATEGY (NNCA-PS)

In this section, we propose to adapt **NNCA** algorithm with some labelled objects to guide the clustering process of the unlabelled objects, i.e., **NNCA** with partial supervision (**NNCA-PS**). The proposed method is divided into two stages. First stage is to select **Np** objects randomly from the dataset to be labelled data ob-
j = \begin{cases} 
\arg \min_{k \in C_i} \sum_{k \in C_i} d_{zk} & \text{if cluster } j \in C_i \\
\arg \min_{k \in C_i} \sum_{k \in C_i} d_{zk} & \text{if cluster } j \notin C_i 
\end{cases} 
\text{if } i \in C_i 
\text{else } i \notin C_i 
\text{if } j \in C_i 
\text{else } j \notin C_i 
\text{if } q \geq 1 
\text{else } q = 1.5 
\text{is used.}

4 DATASETS

Two different types of real-world data are used to investigate whether the proposed algorithm scales well with the size and dimension of the dataset.

4.1 Retinal Images

For performance evaluation, a publicly available dataset is used (STARE, ). The dataset consists of 20 images which are digitised slides captured by a TopCon TRV-50 fundus camera at 35° FOV. Each slide was digitized to produce a 605x700 pixels image, standard RGB, 8 bits per colour channel. Every image has been manually segmented by two observers to produce ground truth vessels segmentation. Ten of these images contain pathology and the other ten are normal, giving a good opportunity to test the proposed method in both normal and abnormal retinas.
Figure 2: (a) Original images, (b) output from the NNCA (hard decision), (c) output from the NNCA-PS (hard decision), and (d) output from the KNN classifier (hard decision).

Algorithm 2 NNCA with partial supervision strategy (NNCA-PS)

- **Step 1: Clustering using NNCA algorithm**
  1. Randomly select $N_p$ points from the ground truth to be labelled objects.
  2. Cluster the $N_p$ objects into $N_C$ cluster using NNCA clustering algorithm.

- **Step 2: Apply the supervision strategy as follow:**
  1. Classify the clusters obtained by NNCA algorithm to the class of its most labelled objects.
  2. For each labelled object, if its cluster is classified to different class (label), then this object will be assigned to the cluster that has the nearest objects and with the same label of it.
  3. Each unlabelled object is assigned to the cluster that has the nearest objects and then classified to the class (label) of this cluster.

5 EXPERIMENTAL RESULTS

5.1 Retinal Images

In our experiments, retinal blood vessels are segmented using the NNCA with partial supervised strategy (NNCA-PS). The performance is measured by the true and false positive rates. These rates are defined in the same way as in (Hoover et al., 2000), where the true (false) positive is any pixel which was hand-labelled as a vessel (not vessel), whose intensity after segmentation is above a given threshold. The true (false) positive rate is established by the dividing the number of true (false) positives by the total number of pixels hand-labelled as vessels (not vessels).

For purposes of comparison, we have compared the performance of NNCA-PS with KNN classifier (Salem and Nandi, 2006a) and RACAL with partial supervision strategy (Salem et al., 2007). For the KNN classifiers, two sets are required; one for training and the other for testing, so the dataset is randomly divided into two sets of images, each contains 5 normal and 5 abnormal images. The training set contains large number of training samples (423500 pixels/image), which is huge and is the main problem with this type of classifiers. To overcome such a problem, a random number of pixels are chosen from the field of view (FOV) of each image in the training set. The targets for these training samples are available from the manually segmented images. The testing

4.2 Breast Cancer Data

Two Wisconsin breast cancer datasets (UCI) are considered in this paper. The first dataset contains 569 samples of 30 features each, and two classes: Benign (class 1 and 357 samples) and Malignant (class 2 and 212 samples). The second dataset contains 683 samples of 9 features each, and two classes: Benign (class 1 and 444 samples) and Malignant (class 2 and 239 samples).
Table 1: NNCA-PS, RACAL and KNN hard decision results (average from 10 images (testing set)).

<table>
<thead>
<tr>
<th>Image type</th>
<th>NNCA-PS</th>
<th>RACAL (Salem et al., 2007)</th>
<th>KNN (Salem and Nandi, 2006a)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Specificity</td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td>Normal</td>
<td>95.4%</td>
<td>90.2%</td>
<td>97.2%</td>
</tr>
<tr>
<td>Abnormal</td>
<td>94.4%</td>
<td>87.8%</td>
<td>96.9%</td>
</tr>
<tr>
<td>All images</td>
<td>94.8%</td>
<td>89.0%</td>
<td>97.0%</td>
</tr>
</tbody>
</table>

Table 2: Average sensitivity at certain specificity values for 10 images.

<table>
<thead>
<tr>
<th>Image type</th>
<th>NNCA-PS</th>
<th>RACAL (Salem et al., 2007)</th>
<th>KNN (Salem and Nandi, 2006a)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity</td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td>Normal</td>
<td>90.8%</td>
<td>85.3%</td>
<td>92.9%</td>
</tr>
<tr>
<td>Abnormal</td>
<td>86.7%</td>
<td>81.0%</td>
<td>93.5%</td>
</tr>
<tr>
<td>All images</td>
<td>88.8%</td>
<td>83.2%</td>
<td>93.2%</td>
</tr>
<tr>
<td>Normal</td>
<td>95.1%</td>
<td>94.1%</td>
<td>91.9%</td>
</tr>
<tr>
<td>Abnormal</td>
<td>92.8%</td>
<td>97.7%</td>
<td>96.6%</td>
</tr>
<tr>
<td>All images</td>
<td>93.9%</td>
<td>95.9%</td>
<td>95.1%</td>
</tr>
</tbody>
</table>

For purposes of comparison, a series of experiments were carried out to examine the performance of NNCA when applying the proposed supervision strategy (NNCA-PS) on breast cancer datasets, where the classification results obtained by NNCA with the supervision strategy on breast cancer dataset 1 are compared with the results of different classifiers (PCA/MDC “Principal Component Analysis / Minimum Distance Classifier” (Theodoridis and Kourtoubas, 2003; Cios et al., 1998), FLDA/MDC “Fisher Linear Discriminant Analysis / MDC” (Cios et al., 1998), MLP “Multi-Layer Perceptron” (Duha et al., 2001), SVM “Support Vector Machine” (Hsu and Lin, 2002), and GP/MDC “Genetic Programming/ MDC” (Guo and Nandi, 2006; Kishore et al., 2002)). In order to achieve fair comparisons as in (Guo and Nandi, 2006), we randomly selected, without replacement, 100 samples (from the entire dataset) for training, and 100 samples for test-
Table 3: Comparison of classification accuracy (%) for breast cancer dataset 1 (testing set) using NNCA-PS and different classifiers, based on 100 experiments.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Best (%)</th>
<th>Average (%)</th>
<th>Std (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA/MDC (Guo and Nandi, 2006)</td>
<td>88.7</td>
<td>88.6</td>
<td>N/A</td>
</tr>
<tr>
<td>FLDA/MDC (Guo and Nandi, 2006)</td>
<td>88.9</td>
<td>88.6</td>
<td>N/A</td>
</tr>
<tr>
<td>MLP (Guo and Nandi, 2006)</td>
<td>97.3</td>
<td>96.2</td>
<td>1.7</td>
</tr>
<tr>
<td>SVM (Guo and Nandi, 2006)</td>
<td>96.7</td>
<td>96.3</td>
<td>0.8</td>
</tr>
<tr>
<td>GP/MDC (Guo and Nandi, 2006)</td>
<td>98.9</td>
<td>97.4</td>
<td>1.5</td>
</tr>
<tr>
<td>NNCA-PS</td>
<td>99.5</td>
<td>97.2</td>
<td>1.2</td>
</tr>
</tbody>
</table>

In order to reduce the amount of a priori knowledge, a small number of objects from the entire dataset are used as labelled objects. In these experiments, the effect of the number of labelled objects on the classification accuracy are investigated. We randomly selected a fraction from the entire dataset to be labelled objects. For each fraction, this process is repeated one hundred times without replacement. The best, average, and standard deviation of classification accuracy are obtained over one hundred runs for each fraction of labelled objects. For breast cancer dataset 1, as demonstrated in Table 4, the best and average classification accuracies increase with the increasing fraction of the labelled objects. As shown, the best and average classification accuracy of 98.2% and 96.3% respectively were achieved at 30% labelled objects, with the lowest being 96.2% and 91.5% for best and average accuracies respectively at 5% labelled objects. By examining the average and standard deviation of the classification performance, when 5% of the entire dataset are labelled, the average performance is the lowest, while it has the highest standard deviation compared with the other fractions of labelled objects. For breast cancer dataset 2 as recorded in Table 5, the standard deviations is lower than the standard deviations of breast cancer dataset 1. It is conjectured that the clusters on breast cancer dataset 2 are more compact with those in breast cancer dataset 1, as indicated in (Salem and Nandi, 2005). For 5% labelled objects and higher, the best classification accuracy is higher than 98% with a small decrease in the standard deviation and a significant increase in the average classification accuracy as demonstrated in Table 5.

When comparing the proposed NNCA-PS with RACAL for breast cancer data classification, where a small number of objects from the entire dataset are used as labelled objects. The average classification accuracy for breast cancer dataset 1 using NNCA-PS is 1% higher than RACAL algorithm while it achieves comparable accuracy for breast cancer dataset 2 as demonstrated in Tables 6 and 7. Moreover, the standard deviation of the classification performance of NNCA-PS for breast cancer dataset 1 is lower than RACAL which favors compact clusters.
Table 6: Comparison of classification accuracy (%) for breast cancer dataset 1 (entire dataset) using NNCA and RACAL with partial supervisions, based on 100 experiments.

<table>
<thead>
<tr>
<th>labelled objects %</th>
<th>NNCA-PS Average(%) ± Std(%)</th>
<th>RACAL (Salem and Nandi, 2006b) Average(%) ± Std(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>91.5 ± 2.3</td>
<td>90.6 ± 4.7</td>
</tr>
<tr>
<td>10</td>
<td>93.1 ± 1.8</td>
<td>92.1 ± 3.2</td>
</tr>
<tr>
<td>15</td>
<td>94.4 ± 1.3</td>
<td>93.5 ± 2.3</td>
</tr>
<tr>
<td>20</td>
<td>95.3 ± 1.0</td>
<td>94.4 ± 1.8</td>
</tr>
<tr>
<td>25</td>
<td>95.6 ± 0.9</td>
<td>94.9 ± 1.6</td>
</tr>
<tr>
<td>30</td>
<td>96.3 ± 0.7</td>
<td>95.2 ± 1.7</td>
</tr>
</tbody>
</table>

Table 7: Comparison of classification accuracy (%) for breast cancer dataset 2 (entire dataset) using NNCA and RACAL with partial supervisions, based on 100 experiments.

<table>
<thead>
<tr>
<th>labelled objects %</th>
<th>NNCA-PS Average(%) ± Std(%)</th>
<th>RACAL (Salem and Nandi, 2006b) Average(%) ± Std(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>98.0 ± 1.2</td>
<td>97.5 ± 1.4</td>
</tr>
<tr>
<td>10</td>
<td>98.1 ± 1.1</td>
<td>97.9 ± 0.3</td>
</tr>
<tr>
<td>15</td>
<td>98.5 ± 0.9</td>
<td>98.2 ± 0.3</td>
</tr>
<tr>
<td>20</td>
<td>98.7 ± 0.8</td>
<td>98.6 ± 0.3</td>
</tr>
<tr>
<td>25</td>
<td>98.7 ± 0.7</td>
<td>98.6 ± 0.3</td>
</tr>
<tr>
<td>30</td>
<td>99.2 ± 0.5</td>
<td>98.6 ± 0.3</td>
</tr>
</tbody>
</table>

while it achieves slightly higher standard deviations in breast cancer dataset 2. This may be the result of the NNCA-PS achieving clustering without any control of cluster sizes while RACAL is constrained with a radius parameter $\delta_0$ which controls the size of the clusters.

6 CONCLUSIONS

In this paper, we have proposed a partial supervision strategy for a recently developed clustering algorithm (NNCA) to act as a classifier. We examined its applicability and reliability using datasets from real-world problems. As shown, the proposed NNCA-PS has the ability to classify pixels of retinal images into those belonging to blood vessels and others not belonging to blood vessels, and it also has the ability to classify breast tumors into either benign or malignant. Experimental results show that the proposed algorithm offers better classification accuracies compared with certain other classifiers.

REFERENCES


STARE. The STARE project. [http://www.ces.clemson.edu/ahoover/stare](http://www.ces.clemson.edu/ahoover/stare).

