An investigation on scaling parameter and distance metrics in semi-supervised Fuzzy c-means

Daphne Teck Ching Lai and Jonathan M. Garibaldi
School of Computer Science, University of Nottingham
Email: dtl@cs.nott.ac.uk and jmg@cs.nott.ac.uk

Abstract—The scaling parameter $\alpha$ helps maintain a balance between supervised and unsupervised learning in semi-supervised Fuzzy c-Means (ssFCM). In this study, we investigated the effects of different $\alpha$ values, 0.1, 0.5, 1 and 10 in Pedrycz and Waletzky’s ssFCM with various amounts of labelled data, 10%, 20%, 30%, 40%, 50% and 60% and three distance metrics, Euclidean, Mahalanobis and kernel-based on the Nottingham Tenovus Breast Cancer dataset and five popular UCI datasets. Higher $\alpha$ values were found to produce better accuracy using Euclidean distance on four datasets out of the six datasets. For Mahalanobis distance, increasing $\alpha$ to improve accuracy is effective up to $\alpha = 1$ and not at $\alpha = 10$ in three out of six datasets. For kernel-based distance, accuracy tend to decrease with increasing $\alpha$ value, which has been observed in four out of six datasets. Such trends in the effects of $\alpha$ values on the classification results using different distance metrics and datasets can be established to form a guide in the selection of $\alpha$. Care should be taken in selection of $\alpha$ value as they are dependant on the distance metric, particularly the Mahalanobis and kernel-based distance metrics, and the dataset used.

I. INTRODUCTION

Fuzzy c-means is a clustering method which allow a data pattern to belong to more than one cluster. This gives a more realistic representation of data than a binary approach as a data pattern may share qualities possessed by more than one cluster. Membership values are used to indicate the degree of belongingness a data pattern has to clusters and thus determine which cluster a data pattern is assigned to. For a data pattern, membership values to each cluster can range from zero to one and the sum of membership values for all clusters must equal to one. A high membership value to a cluster means a high possibility of belonging to this cluster. Semi-supervised Fuzzy c-Means (ssFCM) use some labelled data patterns in the dataset to guide the identification of similar data patterns and to allocate them into meaningful groups. This can be very valuable because labelled data patterns are often sparse and they are time-consuming and labour-intensive to collect.

Pedrycz and Waletzky [1] introduced the following ssFCM objective function containing unsupervised learning in the first term and supervised learning in the second term:

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^p d_{ik}^2 + \alpha \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik} - f_{ik}b_k)^p d_{ik}^2, \quad (1)$$

where $u_{ik}$ is the membership value of data pattern $k$ in cluster $i$, $d_{ik}$ the distance between data pattern $k$ and cluster centre $v_i$, $f_{ik}$ the membership value of labelled data pattern $k$ in cluster $i$, $b_k$ indicates if data pattern $k$ is labelled, $c$ is the number of clusters, $N$ the number of data patterns in the dataset and $p$ is the fuzzifier parameter (which is commonly 2) and $\alpha$ is a scaling parameter for maintaining balance between the supervised and unsupervised learning components. The authors recommended $\alpha$ to be proportional to $N/M$ where $M$ is the number of labelled data.

In this work, we are concerned with the scaling parameter $\alpha$ in Pedrycz and Waletzky’s ssFCM algorithm (Pedrycz97). Previously, we investigated on distance metrics in Pedrycz97 for classifying the Nottingham Tenovus Breast Cancer (NTBC) dataset [2]. We found that Mahalanobis distance performed less favourable than Euclidean distance in Pedrycz97. This was found to be unexpected as the Mahalanobis distance generates hyperelliptical clusters, including spherical ones which Euclidean distance. We expected the Mahalanobis distance to generate clusters that would better fit the dataset. As a continuation of our investigation in NTBC classification using Pedrycz97, we are primarily interested in finding out the effects of the scaling parameter $\alpha$ on the classification results with different distance metrics on the NTBC dataset, particularly if certain values of $\alpha$ can improve Pedrycz97 with Mahalanobis distance. We also wish to address the following questions:

1) How will $\alpha$ affect Pedrycz97 with respect to the distance metrics?
2) Will the effects vary with different datasets?

Our interest in the scaling parameter is motivated by several analysis conducted on the scaling parameter $\alpha$ in other versions of ssFCM. In [3], Bouchachia and Pedrycz found that the number of misclassification decreases when $\alpha$ increases in their ssFCM algorithm with evolving membership. A higher value of $\alpha$ is viewed as an indicator of higher confidence on the goodness of labelled data. Furthermore, it was shown that with higher $\alpha$ values, the clusters are better separated. In another study [4], they showed how higher $\alpha$ values improves clustering results as purity (ratio of the highest number of data patterns having the same labels to the total number of data patterns in that cluster) increases and entropy (distribution of labels in a cluster) decreases on three datasets, Diabetes, Wine and Cancer. A high purity value indicates more data patterns of the same label within each cluster and low entropy value indicate that each cluster produces a good split as it contains exclusively data patterns of one particular label in comparison with data patterns of other labels within that cluster. Also,
they also showed the different effects in noise detection of the algorithm with varying values of $\alpha$ and using several distance metrics. However, high values of $\alpha$ are not always favourable. In [5], Gao and Wu presented that $\alpha$ values in the range [0.05, 0.2] gave the best clustering accuracy on Iris dataset using their pairwise-constrained ssFCM. Values above 0.2 gave less favourable results. Wang et al. [6] expressed challenges in selecting suitable $\alpha$ values for their kernel-based pairwise-constrained ssFCM for different datasets.

From existing studies, we established that high values of $\alpha$ can have different effects on clustering results for different ssFCMs. Thus, it is of great interest to us as to how it will affect the results of Pedrycz97 using different distance metrics for NTBC dataset and further ascertain if these effects are prevalent in other popular datasets. By knowing these effects, we can make better selection of $\alpha$ values to improve clustering results, taking into account of the distance metric and dataset used.

The paper is organised as follows: In Section II, the Pedrycz97 algorithm is explained. The three distance metrics, Euclidean, Mahalanobis and kernel-based, are described in Section III. Next, the experiments carried out are elaborated in Section IV, followed by the presentation of results in Section V. The results are discussed further detail in Section VI. Lastly, we draw conclusions about our finding in Section VII.

II. SEMI-SUPERVISED FUZZY C-MEANS

The objective function of semi-supervised Fuzzy c-means is an extension of FCM’s to include supervised learning component represented as the second term in (1).

The algorithm involves iteratively calculating the cluster centres and the membership matrix $U$ containing $u_{ik}$ to minimise the objective function until a termination criterion is satisfied. In this work, we use ssFCM by Pedrycz and Waletzky [1] because it has been shown to produce good classification results. The algorithm is summarised as follows:

1) Initialise labelled data membership matrix $F$ and initial membership matrix $U^0$
2) Calculate cluster centres $V = [v_i]$ with $U$ using equation:
   \[ v_i = \frac{\sum_{k=1}^{N} u_{ik}^2 x_k}{\sum_{k=1}^{N} u_{ik}^2} \]  
3) Update partition matrix, $U$ using equation:
   \[ u_{ij} = \frac{1}{1 + \alpha} \left( 1 + \alpha (1 - b_j) \sum_{l=1}^{c} f_{ij} + \alpha f_{ij} b_j \right) \frac{\sum_{k=1}^{N} (d_{ij}^2)^2}{\sum_{k=1}^{N} d_{ik}^2} \]
4) If $\|U' - U\| < \epsilon$, stop. Else, go to step 2 with $U = U'$

III. DISTANCE METRICS

In this section, we briefly describe the distance metrics used. These distance metrics, their differences and behaviours in ssFCM are discussed in detail in [4].

A. Mahalanobis

The Mahalanobis distance is formally defined [7] as:
\[ d_M(x) = \sqrt{(x - \mu)^T S^{-1}(x - \mu)} \]  

B. Euclidean

The Euclidean distance metric forms spherical clusters and does not reflect scale differences among dimensions in high-dimensional datasets. It is computed as follows:
\[ d_E^2(i, k) = ||x_k - v_i||^2 \]  

C. Kernel-based

The kernel methods solve non-linear problems by mapping the input space into higher dimensional space (the ‘kernel trick’ [8]), which is applied to distances metrics in [9]. The idea here is to transform $x_k$, a data point from a $D$-dimensional input space to a higher $F$-dimensional space resulting in $\Phi(x_k)$. The kernel-based distance is defined as:
\[ d_K^2(i, k) = ||\Phi(x_k) - \Phi(v_i)||^2 \]

We use Gaussian radial basis function as the kernel function in the form:
\[ K(a, b) = e^{-\frac{||a - b||^2}{2\sigma^2}} \]

This yield a distance of the form
\[ d_K^2(i, k) = 2(1 - K(x_k, v_i)) \]
IV. EXPERIMENTS

A. Dataset

The Nottingham Tenovus Breast Cancer dataset contains 25 immunohistochemical features for 1076 patients. There are three main clinical groups, Luminal, Basal and HER2 and six subgroups where class 1, 2 and 3 belongs to the Luminal group, class 4 and 5 to the Basal group and class 6 to HER2. Each feature is not normally distributed and they have been scaled to a range between 0 and 300. The UCI datasets [10] used are Ionosphere (I), Page Blocks (PB), Pima Indians Diabetes (PID), Wine and Wisconsin Original Breast Cancer (WOBC). In WOBC, the attribute with missing values have been removed. Table I shows the characteristics of the four datasets.

B. Experimental Design and Set-up

Figure 1 illustrates the way the experiment is conducted. Provided labels are used to generate membership values used to initialise the supervision matrix $F$. Instead of using random initialisation of membership values, we use the supervision matrix $F$ to initialise the membership matrix $U^0$. In doing so, a better starting point is given to the algorithm instead of a random starting point. For the NTBC dataset, we use only the 663 data patterns classified by Soria et al. for investigation and the 413 data patterns which are not classified are disregarded.

We experimented with three distance metrics, Euclidean, Mahalanobis and kernel-based, varying amounts of labelled data: 10%, 20%, 30%, 40%, 50% and 60% of the dataset and with varying values of $\alpha$: 0.1, 0.5, 1 and 10. We run the Pedrycz97 algorithm with each varying amount across 100 different sets of labelled data. To select data patterns to be labelled, an equal number of data patterns is randomly selected from each class. This is to prevent data patterns from a particular class to be selected more frequently than others. From the selected labelled data, we initialise the $F$ matrix. To initialise membership values in $F$, the selected labelled data patterns belonging to their respective clusters will be given a membership of 0.9 and (1-0.9)/(c-1) for the remaining clusters. The high 0.9 membership value is arbitrarily chosen to indicate a data pattern’s high possibility of belonging to the cluster while the much lower membership value in the remaining clusters indicates otherwise. Unlabelled data patterns have a membership value of $1/c$ to indicate equal possibility of belonging to the clusters.

In the original Pedrycz97, all data patterns are assigned memberships based on their given labels and stored in $F$. They are then selected to be labelled and unlabelled for the algorithm using the boolean vector in Equation 1, b. In our case, we have selected the labelled data for the algorithm and generated their memberships prior to running the algorithm. We set our $F = U^0$, where they contain memberships of both labelled and unlabelled data and $b_k$ is 1 for all $k$ (in Eq. 3).

To determine the cluster of a data pattern $x_k$, we choose the cluster with the highest membership value. We used numerical labels such that cluster 1 is class 1, cluster 2 is class 2 and so forth. The classes assigned by ssFCM are then compared with the provided labels and evaluated using various evaluation techniques, which will be explained next.

C. Evaluation and Analysis

To evaluate the accuracy of the algorithm in clustering the dataset, we calculate the average number of agreeing results with Soria’s classification over the total number of data patterns across the 100 runs.

To study the effects of $\alpha$, we look at the accuracy achieved using different distance metrics on different data sets as well as the correlation between accuracy and $\alpha$ for each distance metric. The correlation is based on Pearson’s correlation which can be calculated by dividing the covariance of two variables by the product of their respective standard deviations. The higher the correlation coefficient, the stronger the correlation between accuracy and $\alpha$, which can indicate whether if the effects of $\alpha$ on Pedrycz97 is a positive or negative one.

V. RESULTS

Figure 2 shows graphs of accuracy against percentage of labelled data for the six datasets with three distance metrics and $\alpha$ values of 0.1, 0.5, 1 and 10.

In NTBC, accuracy increases with $\alpha$ and percentage of labelled data for Mahalanobis distance but decreases with $\alpha$ for kernel-based distance. For Euclidean distance, slight increase with $\alpha$ is found at 10% labelled data.

In Ionosphere, accuracy increases with $\alpha$ and percentage of labelled data for Euclidean and kernel-based distance but, at $\alpha = 10, 50%$ and 60% labelled data, accuracy decrease.

In Page Blocks, accuracy increases with increasing values of $\alpha$ for Mahalanobis and Euclidean distances. Higher amount of labelled data do not improve results except at $\alpha = 10$ for

<table>
<thead>
<tr>
<th>Dataset</th>
<th>N</th>
<th>c</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nottingham Tenovus Breast Cancer (NTBC)</td>
<td>1076</td>
<td>6</td>
<td>25</td>
</tr>
<tr>
<td>Ionosphere (I)</td>
<td>351</td>
<td>2</td>
<td>33</td>
</tr>
<tr>
<td>Page Blocks (PB)</td>
<td>5473</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Pima Indians Diabetes (PID)</td>
<td>768</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Wine</td>
<td>178</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>Wisconsin Original Breast Cancer (WOBC)</td>
<td>699</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>
Fig. 2. Classification results on the datasets using 3 different distance metrics and 4 scaling parameter values.
Euclidean distance. And, for kernel-based distance, accuracy is decreasing as $\alpha$ increases.

In PID, accuracy increases with $\alpha$ and percentage of labelled data for Euclidean and kernel-based distance. At $\alpha = 0.1$, increased amounts of labelled data improved accuracy very slightly. For Mahalanobis distance, accuracy drops slightly at $\alpha = 10$.

In Wine, accuracy appears to be similar with increasing $\alpha$ for Mahalanobis distance. For Euclidean distance, accuracy increases with $\alpha$ and percentage of labelled data. But, accuracy decreases with $\alpha$ for kernel-based distance.

In WOBC, accuracy increases with $\alpha$ and percentage of labelled data for Mahalanobis distance but remains the same for Euclidean distance and kernel-based distance.

In Table II, the correlation between accuracy and $\alpha$ for each distance metric is shown. We can observe stronger correlation in PB, PID, Wine and WOBC using Euclidean distance than Mahalanobis distance. There is negative correlation for NTBC, PB, Wine and WOBC using kernel-based distance. Comparing the results of this table with Figure 2, positive correlation suggests an overall positive contribution from $\alpha$ to the results while negative correlation suggests otherwise. Positive correlation indicates how much increase in $\alpha$ is related to increase in accuracy as observed in PID with Mahalanobis distance where accuracy is slightly lower at $\alpha = 10$ than at $\alpha = 1$. The slightly lower accuracy at a higher $\alpha$ value will decrease the correlation even though accuracy is increasing with $\alpha$ at 0.1, 0.5 and 1 in Figure 2.

VI. DISCUSSION

For the NTBC dataset, the most improved results was found at $\alpha = 1$ for Mahalanobis distance and no further improvement was observed at $\alpha = 10$. However, better accuracy results was produced using Euclidean distance than Mahalanobis distance.

Higher $\alpha$ values increases contribution from labelled data and this has been observed in results of Figure 2 on Ionosphere, PB, PID and Wine for Euclidean distance, on NTBC, PB, PID and WOBC datasets for Mahalanobis distance and on PID for kernel-based distance. These results are also reflected in Table II when the strong correlation with values above 0.5 between accuracy and $\alpha$ suggest overall positive effects from $\alpha$ on accuracy. These results and analysis are evidence to suggest higher $\alpha$ values improves accuracy, but not for all distance metrics and datasets.

Where very high accuracy levels have already been achieved at low percentage of labelled data, increasing $\alpha$ values will not further improve results as can be observed in WOBC for Euclidean distance in Figure 2.

Despite the overall positive contribution from $\alpha$ with a particular distance metric, higher accuracy is found using distance metrics less affected by $\alpha$ values. For instance, Mahalanobis distance in NTBC and Euclidean distance in Ionosphere both improved accuracy with increased $\alpha$ but Euclidean distance and Mahalanobis distance achieved better accuracy in the respective datasets.

For Mahalanobis distance, if increase in accuracy is found on a dataset, it stops at $\alpha = 1$ as similar accuracy results is found with $\alpha = 10$, observed in three out of six datasets. This is also reflected in Table II where weak correlation between accuracy and $\alpha$ suggests that the increase in $\alpha$ between values of 1 and 10 is not bringing a great overall increase in accuracy, as shown in Figure 2.

For kernel-based distance, higher $\alpha$ values tend to decrease accuracy, as observed in NTBC, Ionosphere, PB and Wine. For Ionosphere and WOBC, accuracy decrease drastically at $\alpha = 10$ and 60% labelled data.

VII. CONCLUSION

The effects of scaling parameter $\alpha$ on Pedrycz97 is dependent on the distance metric and the dataset. With increased $\alpha$ values, positive effects are mostly found in Pedrycz97 using Euclidean distance, on Ionosphere, PB, PID, Wine and WOBC datasets although this conclusion cannot be generalised to the NTBC dataset, nor to other datasets. Despite the positive effects from increased $\alpha$ values on Euclidean distance, Mahalanobis distance, which is less affected by increased $\alpha$ values, achieves higher accuracy on four out of the six datasets.

Some other trends were observed such as the tendency to decrease in accuracy using kernel-based distance with increasing $\alpha$ values and the small increase in accuracy using Mahalanobis distance tend to stop at $\alpha = 1$.

Unfortunately, there is no strong trends found from our investigation to provide a definitive guide on the selection of $\alpha$ for the three different distance metrics on all datasets. Rather, a form of informal guide can be established. We found that with Euclidean and Mahalanobis distance, increasing $\alpha$ can bring positive effects to accuracy but not to all datasets. For Mahalanobis, this is effective at $\alpha = 1$ and not $\alpha = 10$. For kernel-based distance, increasing $\alpha$ tend to have a detrimental effect to accuracy. The knowledge of these effects should better prepare us in future investigations when handling with $\alpha$ values. A possible approach to ascertain we obtain the best results is to run the algorithm across various distance metrics and $\alpha$ values, validate their performance using several evaluation techniques and cast a vote based on the evaluations for the distance metric and $\alpha$ value that produces the best result.

REFERENCES


