Chapter 6

Dynamically Determining the Number of Clusters Found by a Local Network Neighbourhood Artificial Immune System

A challenge in data clustering is to determine the optimal number of clusters in a data set. Section 2.4 discussed a number of approaches to validate and determine the number of clusters in a data set. These approaches include validation of the formed clusters by visual inspection and/or multiple execution of the clustering algorithm, each time with a different number of clusters and validating the clustered data set with a cluster validity index. The former visual approach becomes infeasible for multidimensional problems where the number of dimensions is greater than three and even though the latter multiple execution approach is familiar in the field, it is computationally expensive and time consuming. Therefore a clustering technique or model which can dynamically determine the number of clusters in a data set and which is computationally inexpensive will have an added advantage.

Although most of the existing network based artificial immune models do not require any user specified parameter of the number of required clusters to cluster the data, these models do have a drawback in the techniques used to determine the number of clusters. These techniques and their drawbacks were discussed in section 5.5.6. All of the techniques share a mutual drawback which is the user specified parameter of the number of required clusters.

This chapter discusses some of the existing data clustering methods to dynamically determine the number of clusters in a data set. Two techniques are then proposed which can be used with the local network neighbourhood artificial immune model to dynamically determine the number of clusters in a data set. The first technique utilises cluster validity indices and is similar to the multiple execution approach, though computationally less expensive. The second technique is based on sequential deviation outlier detection, which was discussed in section 2.6. The end result of both techniques is an enhanced LNNAIS model that can dynamically determine the number of clusters in a data set.

Experimental results of K-means clustering using the multiple execution technique are compared with the results of the proposed LNNAIS techniques.

6.1 Dynamic Data Clustering Methods

Dynamically determining the optimal number of clusters in a data set is a challenging task, since *a priori* knowledge of the data is required and not always available. As discussed in section 2.4, cluster validity indices can be used with a multiple execution of the clustering algorithm to dynamically determine the number of clusters. A disadvantage of the multiple execution approach is that the technique is computationally expensive and time consuming. Other techniques and clustering models have also been proposed in the literature and are discussed next.

Ball and Hall [10] proposed the Iterative Self-Organising Data Analysis Technique (ISODATA) to dynamically determine the number of clusters in a data set. As with K-means clustering, ISODATA iteratively assigns patterns to the closest centroids. Different to K-means clustering, ISODATA utilises two user-specified thresholds to respectively merge two clusters (if the distance between their centroids is below the first threshold) and also split a cluster into two clusters (based on the second threshold). Even though ISODATA has an advantage above K-means clustering to dynamically determine the number of clusters in the data set, ISODATA has two additional user parameters (merging and splitting thresholds) which have an effect on the number of clusters determined. A similar model to ISODATA is the Dynamic Optimal Cluster-seek (DYNOC) which was proposed by Tou [172]. DYNOC also follows an iterative approach with splitting and merging of clusters but at the same time maximises the ratio of the minimum inter-clustering to the maximum intra-clustering distance. DYNOC also requires a user specified parameter which determines the splitting of a cluster. SYNERACT was proposed by Huang [87] as an alternative to ISODATA. SYNERACT uses a hyperplane to split a cluster into smaller clusters for which the centroids need to be calculated. Similar to ISODATA and DYNOC, an iterative approach is followed to assign patterns to available clusters. Even though SYNERACT is faster than ISODATA and does not require the initial location of centroids or the number of clusters to be specified, SYNERACT does require values for two parameters which have an effect on the splitting of a cluster.

Veenman proposed a partitional clustering model which minimises a cluster validity index in order to dynamically determine the number of clusters in a data set [175]. The initial number of clusters is equal to the number of patterns in the data set. An iterative approach is followed to determine the splitting and merging of clusters. In each iteration, tests which are based on the minimisation of the cluster validity index determine the splitting or merging of clusters. The proposed algorithm has similar drawbacks as the multiple execution approaches, namely that the model is computationally expensive and has user parameters for the cluster validity index which influences the clustering results.

Another K-means based model was proposed by Pelleg and Moore [128] and uses model selection. The model is called X-means and initially start with a single cluster, $K = 1$ (which is the minimum number of clusters in any data set). The first step is then to apply K-means clustering on the K clusters which are then split in a second step according to a Bayesian Information Criterion (BIC) [106]. If the BIC is improved with the splitting of the clusters, the newly formed clusters are accepted, otherwise it is rejected. These steps are repeated until a user specified upper bound on K is reached. X-means clustering dynamically determines the number of clusters in the data set as the value of K which has the best BIC value. X-means also has a drawback of a user specified parameter for the upper bound on K. Hamerly and Elkan proposed a similar model as X-means clustering, called G-means clustering [72]. G-means also starts with a small value of K but only splits clusters which data do not have a Gaussian distribution. This is also a drawback of G-means clustering, since it is assumed that the data has spherical and/or elliptical clusters [72].

There are also other models proposed in the literature which is either based on K-means clustering or utilises K-means with similar approaches of splitting and merging clusters. These models are *Snob* [176] and Modified Linde-Buzo-Gray (MLBG) [154]. All of the discussed models suffer from either user parameters which influence the clustering results or can only cluster data sets with specific characteristics.

The following section proposes two techniques which can be used with the local network neighbourhood artificial immune model to dynamically determine the number of clusters in a data set.

6.2 Dynamic Clustering Techniques for LNNAIS

This section proposes two alternative techniques which can be used by LNNAIS to dynamically determine the number of clusters in a data set. Both of these techniques have advantages and drawbacks which are also discussed. This section first recapitulates the technique used by LNNAIS to determine a user specified number of clusters (as discussed in section 5.5.6).

Different to other network based AIS models, LNNAIS need not to follow a hybrid approach nor a proximity matrix of network affinities in order to determine the formed ALC networks in the ALC population. This is due to the index based neighbourhood topology utilised by LNNAIS. An index based neighbourhood results in the formation of a ring-like network topology as illustrated in figure 5.3. The required number of ALC networks (or rather clusters), K, can be determined by sorting the network affinities in descending order and selecting the first K network affinities in the sorted set. The K selected network affinities determine the boundaries of the ALC networks.

Figure 5.3 illustrates this technique where $K = 3$. Separate ALC networks are formed by pruning the edges of the K selected boundaries (illustrated as dotted lines in figure 5.3). The centroid of each of the formed ALC networks (illustrated as clouds in figure 5.3) is calculated using equation (2.18). An alternative approach to sorting the network affinities is to plot the network affinities against the numbered edges (as illustrated in figure 6.1). The K edges in the graph with the lowest plotted network affinity (highest Euclidean distance) are then selected as the boundaries of the ALC networks.

Iterative Pruning Technique (IPT): Instead of specifying K, the above pruning technique is done with an iterative value of K. First K is set to 2 where only the top two boundaries are selected for pruning (top two network affinities in the sorted set of network affinities). The quality of the clusters is then measured with a cluster validity index of choice. The same procedure is followed for $K = \{3, 4, 5, \ldots, \mathcal{B}_{max}\}$, measuring the quality with a cluster validity index for each value of K. The value of K with the highest (or lowest, depending on the validity index used)

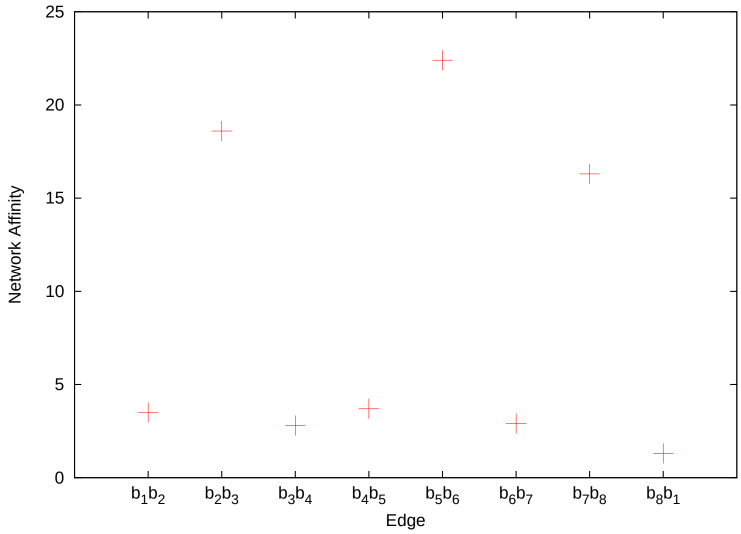

Figure 6.1 Network Affinity Plot

cluster validity index is then selected as the optimal number of clusters. It is also possible to set a minimum and maximum for K, but this can also be seen as a drawback since two parameters need to be specified. If no minimum/maximum is specified it could also be a time consuming task (to a lesser extent when compared to the multiple execution technique) to iterate through all values of K, especially with large values of $\mathcal{B}_{max}$. Whether K is bounded by a minimum/maximum or not, an advantage of the Iterative Pruning Technique to dynamically determine the number of clusters is that the LNNAIS model needs not to be executed for each value of K as in the case of the multiple execution technique. Therefore the Iterative Pruning Technique is computationally less expensive.

Sequential Deviation Outlier Technique (SDOT): Section 2.6 defined outliers and explained three different approaches for outlier detection. One of these approaches is the sequential exception technique which forms part of the deviation based techniques for outlier detection. The reader is referred to section 2.6 for a refresher on the sequential exception technique. As illustrated in figure 6.1, the network affinities which form clear boundaries between the ALC networks tend to be outliers to the remainder of the network affinities.

In the context of dynamically determining the boundaries between the ALCs in LNNAIS, the sequential exception technique can be applied to a sorted set (descending) of network affinities between the ALCs in LNNAIS. The set of network affinities is sorted to guarantee that the lowest network affinities (potential outliers with the highest Euclidean distance) forms part of the first sequential subsets. The first subset, S_1, will then contain the lowest network affinity, followed by S_2 which consists of S_1 and the second lowest network affinity and so forth. The function of dissimilarity $D(S_o)$ in equation (2.67) is calculated as the variance between the network affinities in subset S_o. Therefore the exception set S_e contains the lowest network affinities between the ALCs in LNNAIS and eventually determines the boundaries between the ALCs.

An added advantage of the Sequential Deviation Outlier Technique (SDOT) is that not only is the technique computationally less expensive, but it also has no need for any boundary constraints on K. K is solely determined by the size of S_e. Furthermore, SDOT is a non-parametric technique. The following section discusses the time complexity of SDOT and IPT.

6.3 Time Complexity of SDOT and IPT

The time complexity of both SDOT and IPT are based on the complexity of sorting the network affinities between the ALCs in the ALC population and determining the number of boundaries between the ALCs in the ALC population of size $\mathcal{B}_{max}$. The maximum number of boundaries in an ALC population of size $\mathcal{B}_{max}$ is $\mathcal{B}_{max}$. The time complexity of sorting the $\mathcal{B}_{max}$ network affinities depends on the sorting algorithm used. Assume the time complexity of the sorting algorithm is some constant, χ_1, and that the time complexity of the selected validity index is χ_2. The worst case of time complexity for IPT is when the clustering quality of all possible boundaries needs to be calculated, giving a time complexity of $O(\chi_2 \mathcal{B}_{max} |\mathcal{A}| N)$ where $|\mathcal{A}|$ is the size of the data set that needs to be partitioned and N is the number of dimensions of data set $\mathcal{A}$. The $\mathcal{B}_{max}$ and χ_2 parameters are fixed in advance and usually $\mathcal{B}_{max} << |\mathcal{A}|$. If $\mathcal{B}_{max} << |\mathcal{A}|$ then the time complexity of IPT is $O(|\mathcal{A}|)$ and if $\mathcal{B}_{max} \approx |\mathcal{A}|$ then the time complexity of IPT is $O\left(|\mathcal{A}|^2\right)$. Focusing on SDOT, the maximum number of smoothing factor function evaluations is equal to the size of the ALC population, which is $\mathcal{B}_{max}$. Assume the time complexity of the smoothing function is χ_3. The worst case of time complexity for SDOT is when the smoothing factor of $\mathcal{B}_{max}$ subsets need to be calculated to determine the exception set S_e (as discussed in section 2.6). This gives a time complexity of $O(\chi_3 \mathcal{B}_{max})$ for SDOT. Compared to the time complexity of IPT, the time complexity of SDOT is not influenced by the size of data set $\mathcal{A}$ and also not by the number of dimensions, N.

The following section discusses and compares the results obtained from K-means clustering using the multiple execution technique to determine the number of clusters in a data set and the

Table 6.1 LNNAIS Parameter Values

Data set	$\mathcal{B}_{max}$	ρ	ε_{clone}
iris	25	3	5
two-spiral	20	3	5
hepta	40	3	5
engytime	20	3	10
chainlink	40	3	5
target	30	3	5
ionosphere	20	3	20
glass	20	3	5
image segmentation	30	3	20
spambase	10	5	20

results obtained from LNNAIS using SDOT and IPT to determine the number of clusters in a data set.

6.4 Experimental Results

This section compares and discusses the clustering results obtained by K-means clustering, LNNAIS using IPT, and LNNAIS using SDOT to dynamically determine the number of clusters in a data set. K-means utilises the multiple execution technique with the Q_{DB} (as defined in equation (2.41)) and Q_{RT} (as defined in equation (2.51)) validity indices, referred to as KM$_{DB}$ and KM$_{RT}$, respectively. Two of the LNNAIS models utilises the iterative pruning technique with the same Q_{DB} and Q_{RT} validity indices as K-means, referred to as LNN$_{DB}$ and LNN$_{RT}$, respectively. For the Q_{RT} validity index, parameter c was set to 10 in all the experiments. The value of c was found empirically and values of $c > 10$ have no effect on Q_{RT} for all the data sets. LNN$_{SDOT}$ utilises the sequential deviation outlier technique and thus need no validity index.

All experimental results reported in this section are averages taken over 50 runs, where each run consisted of 1000 iterations of a data set. The parameter values for each data set were empirically found to deliver the best performance for each of the algorithms. The value of K was iterated from $K = 2$ to $K = 12$ for all data sets. Table 6.1 summarises the parameter values used by the respective algorithms for each data set. The clustering quality of the algorithms (based on the number of clusters determined by each of the algorithms) is determined by the Q_{ratio} index, J_{intra} and J_{inter} performance measures (as defined in equations (2.49),(2.17) and (2.16), respec-

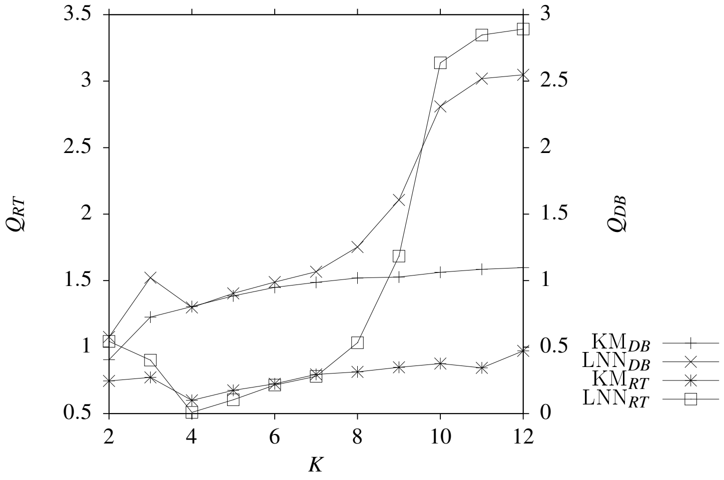

Figure 6.2 Optimal number of clusters obtained by K-means and LNNAIS for the iris data set

tively). The following hypothesis is defined to determine whether there is a difference between the clustering quality of two algorithms for a specific data set or not:

- *Null* hypothesis, H_0: There is no difference in the clustering quality, Q_{ratio}.
- *Alternative* hypothesis, H_1: There is a difference in the clustering quality, Q_{ratio}.

A non-parametric Mann-Whitney U test with a 0.95 confidence interval ($\alpha = 0.05$) was used to test the above hypothesis. The result is statistically significant if the calculated probability (p-value is the probability of H_0 being true) is less than α. In cases where there is a statistical significant difference between the clustering quality of two algorithms, the algorithm with the lowest critical value, z, tends to find clusters in the data set with a higher quality. The results for each of the data sets used are discussed next.

6.4.1 Iris data set

Figure 6.2 illustrates the Q_{RT} values where $c = 10$ for KM$_{RT}$ and LNN$_{RT}$ on the $y1$-axis at different values of K. The Q_{DB} values for KM$_{DB}$ and LNN$_{DB}$ is illustrated on the $y2$-axis of figure 6.2. Figure 6.2 highlights that the optimal number of clusters in the iris data set is obtained by KM$_{RT}$ and LNN$_{RT}$ at $K = 4$ and by KM$_{DB}$ and LNN$_{DB}$ at $K = 2$. Therefore, the optimal range of K is $K = 2$ to $K = 4$ for the iris data set. The average number of clusters determined by LNN$_{SDOT}$ is $K = 2.64$ which falls within the optimal range of K as determined above. Figure 6.3 illustrates for the iris data set the number of clusters respectively determined by the SDOT and IPT

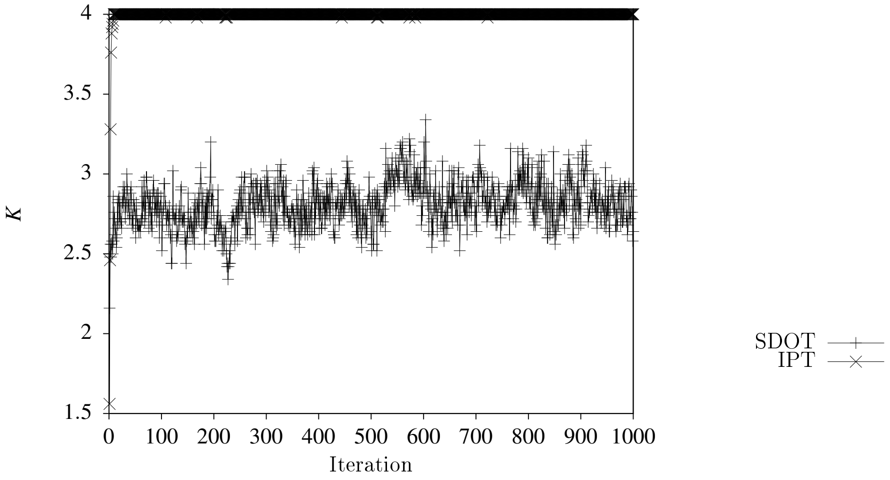

Figure 6.3 Convergence of LNNAIS using SDOT and IPT to optimal K for iris data set

techniques over time. The value of K for IPT rapidly increases to 4 in the first few iterations and remains at 4 for the most of the remaining iterations. The value of K for SDOT increases to 2.7 and oscillates between 2.4 and 3.3 around an average K of 2.64 for the remaining iterations. Since LNNAIS is a stochastic algorithm which utilises a dynamic population of ALCs, the affinities between neighbouring ALCs change over time. Thus, it is expected that the network boundaries detected by SDOT to determine the value of K will also differ over time and oscillate around an average K. Figure 6.4 illustrates a histogram of the frequency distribution of the number of clusters determined by LNN$_{SDOT}$ for the iris data set. The figure illustrates that LNN$_{SDOT}$ has high frequencies at $K = 2$ and $K = 3$. The figure also illustrates that LNN$_{SDOT}$ obtained $K = 4$ for some of the runs, still being within the optimal range of K for the iris data set.

Table 6.2 shows the results obtained by the different models to determine the optimal number of clusters in the iris data set. Referring to table 6.12, the Mann-Whitney U statistical hypothesis test rejects H_0 that the Q_{ratio} means are the same at a 0.05 level of significance between KM$_{RT}$ and LNN$_{SDOT}$ ($z = 7.58$, $p < 0.001$) and between LNN$_{RT}$ and LNN$_{SDOT}$ ($z = 6.69$, $p < 0.001$). Thus, there is a statistical significant difference in the clustering quality, Q_{ratio}, of the iris data set between KM$_{RT}$ and LNN$_{SDOT}$ and between LNN$_{RT}$ and LNN$_{SDOT}$. LNN$_{SDOT}$ tends to find clusters in the iris data set with a higher quality.

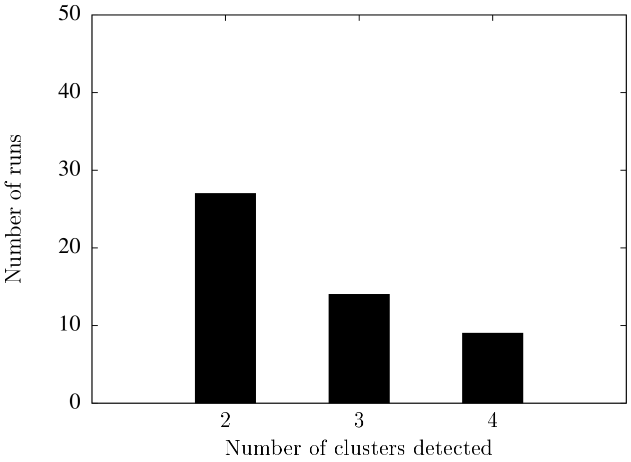

Figure 6.4 Histogram of the number of clusters detected in the iris data set by LNN$_{SDOT}$

Table 6.2 Descriptive Statistics: Iris

Algorithm	K	J_{intra}	J_{inter}	Q_{ratio}	Q_{DB}
KM$_{DB}$	2.00 (± 0.00)	0.856 (± 0.000)	3.927 (± 0.000)	0.218 (± 0.000)	0.405 (± 0.000)
KM$_{RT}$	4.00 (± 0.00)	0.581 (± 0.021)	3.048 (± 0.153)	0.575 (± 0.165)	0.805 (± 0.045)
LNN$_{DB}$	2.00 (± 0.00)	0.923 (± 0.097)	3.994 (± 0.352)	0.233 (± 0.035)	0.432 (± 0.072)
LNN$_{RT}$	4.00 (± 0.00)	0.618 (± 0.036)	3.126 (± 0.221)	0.488 (± 0.154)	0.798 (± 0.154)
LNN$_{SDOT}$	2.64 (± 0.77)	0.788 (± 0.109)	3.738 (± 0.466)	0.364 (± 0.552)	0.643 (± 0.858)

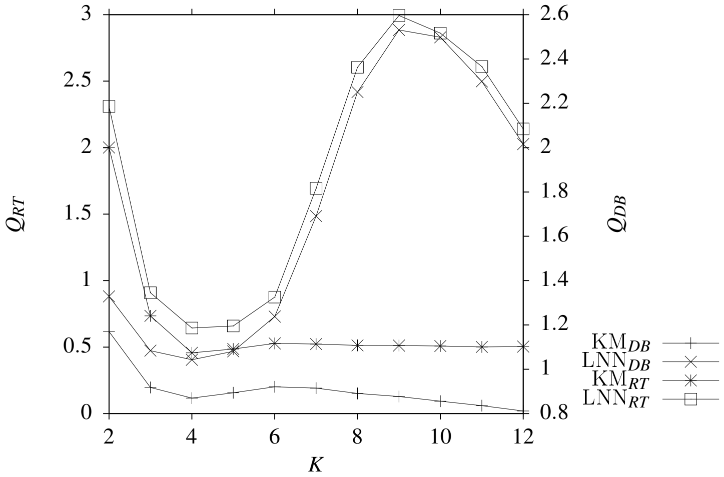

Figure 6.5 Optimal number of clusters obtained by K-means and LNNAIS for the two-spiral data set

6.4.2 Two-spiral data set

The optimal range of K as determined by the different models for the two-spiral data set is $[3, 12]$ (as illustrated in figure 6.5). Furthermore, figure 6.5 shows that although the optimal number of clusters in the two-spiral data set is obtained by KM$_{DB}$ at $K = 12$, the majority of the models obtain the optimal number of clusters in the two-spiral data set at $K = 4$. The average number of clusters determined by LNN$_{SDOT}$ is $K = 4.06$ which is similar to the optimal number of clusters obtained by the majority of the models. Figure 6.6 illustrates a histogram of the frequency distribution of the number of clusters determined by LNN$_{SDOT}$ for the two-spiral data set. The figure illustrates that LNN$_{SDOT}$ has high frequencies for $2 \leq K \leq 5$. Figure 6.7 illustrates that for the two-spiral data set the IPT technique converges to $K = 4$ and SDOT oscillates between $K = 3.5$ and $K = 5$ around an average $K = 4.2$ which is near the value of K as determined by IPT. The statistical hypothesis test rejects H_0 that the Q_{ratio} means are the same between KM$_{RT}$ and LNN$_{SDOT}$ ($z = 8.328$, $p < 0.001$). There is thus a statistical significant difference between the clustering quality of KM$_{RT}$ and LNN$_{SDOT}$. KM$_{RT}$ tends to find clusters in the two-spiral data set with a higher quality than LNN$_{SDOT}$. There is however no statistical significant difference between the Q_{ratio} means of LNN$_{RT}$ and LNN$_{SDOT}$ (statistical hypothesis test accepts H_0, refer to table 6.12). Table 6.3 shows the results obtained by the different models to determine the optimal number of clusters in the two-spiral data set.

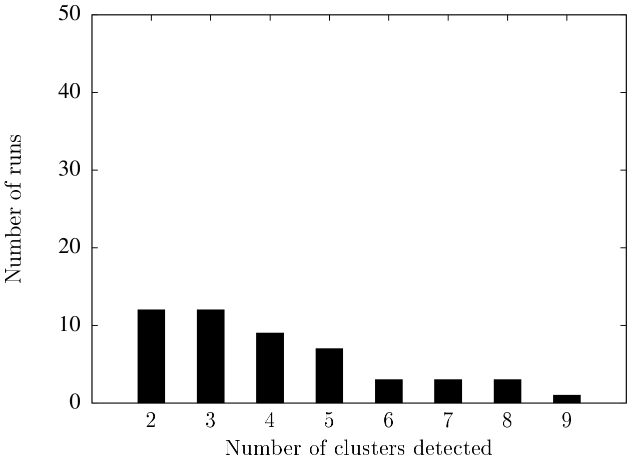

Figure 6.6 Histogram of the number of clusters detected in the two-spiral data set by LNN_{SDOT}

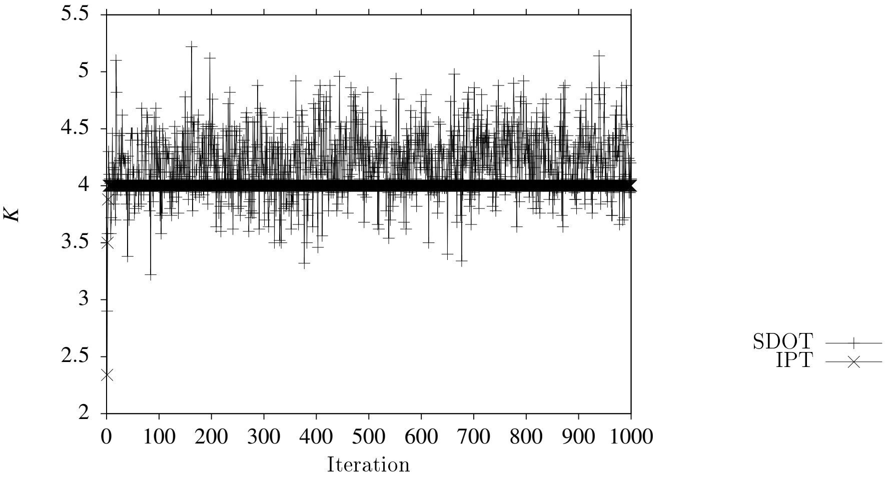

Figure 6.7 Convergence of LNNAIS using SDOT and IPT to optimal K for two-spiral data set

Table 6.3 Descriptive Statistics: Two-spiral

Algorithm	K	J_{intra}	J_{inter}	Q_{ratio}	Q_{DB}
KM$_{DB}$	12.00 (± 0.00)	0.212 (± 0.004)	1.018 (± 0.024)	0.504 (± 0.084)	0.812 (± 0.034)
KM$_{RT}$	4.00 (± 0.00)	0.369 (± 0.003)	0.993 (± 0.011)	0.437 (± 0.016)	0.870 (± 0.031)
LNN$_{DB}$	3.00 (± 0.00)	0.477 (± 0.023)	1.115 (± 0.146)	0.544 (± 0.122)	0.992 (± 0.191)
LNN$_{RT}$	4.00 (± 0.00)	0.405 (± 0.019)	1.021 (± 0.099)	0.616 (± 0.149)	1.043 (± 0.168)
LNN$_{SDOT}$	4.06 (± 1.89)	0.427 (± 0.087)	1.021 (± 0.088)	0.699 (± 0.736)	1.116 (± 0.537)

6.4.3 Hepta data set

The average number of clusters determined by LNN$_{SDOT}$ for the hepta data set is $K = 6.64$ which is close to the true number of clusters in the hepta data set (hepta consists of seven clusters) and falls within the optimal range of K which is $[4,7]$ (as illustrated in figure 6.8). Figure 6.9 illustrates a histogram of the frequency distribution of the number of clusters determined by LNN$_{SDOT}$ for the hepta data set. Figure 6.9 highlights that LNN$_{SDOT}$ has the highest frequency at seven clusters, which is the number of clusters in the hepta data set. Figure 6.10 illustrates for the hepta data set the number of clusters respectively determined by the SDOT and IPT techniques over time. The value of K for IPT converges to 6. The value of K for SDOT oscillates between $K = 6$ and $K = 7$ around an average K of 6.7 for the remaining iterations. Referring to table 6.12, there is a statistical significant difference between the clustering quality of KM$_{RT}$ and LNN$_{SDOT}$ and between LNN$_{RT}$ and LNN$_{SDOT}$. Although KM$_{RT}$ and LNN$_{RT}$ tend to find clusters in the hepta data set with a higher quality than LNN$_{SDOT}$ (refer to table 6.4), LNN$_{SDOT}$ was able to determine the number of clusters in the hepta data set more accurately.

6.4.4 Engytime data set

Table 6.5 shows the results obtained by the different models to determine the optimal number of clusters in the engytime data set. Figure 6.11 illustrates that the optimal range of K for the engytime data set is $2 \leq K \leq 7$ (also shown in table 6.5). LNN$_{SDOT}$ determined the number of clusters in the engytime data set as $K = 3.86$. The histogram of the frequency distribution of the number of clusters determined by LNN$_{SDOT}$ for the engytime data set illustrates that LNN$_{SDOT}$

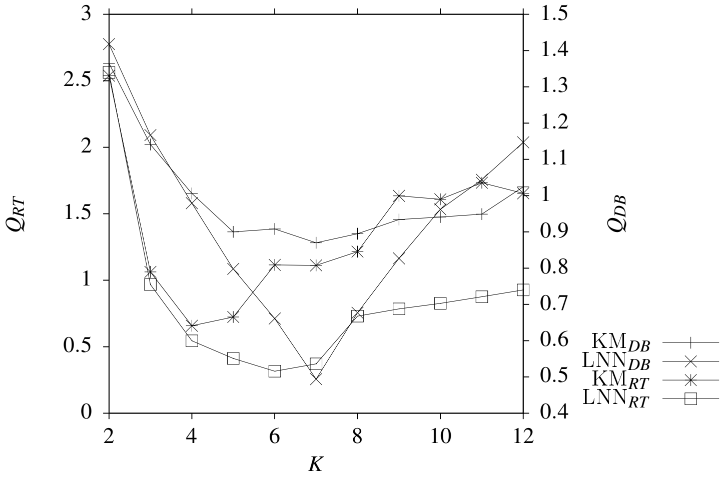

Figure 6.8 Optimal number of clusters obtained by K-means and LNNAIS for the hepta data set

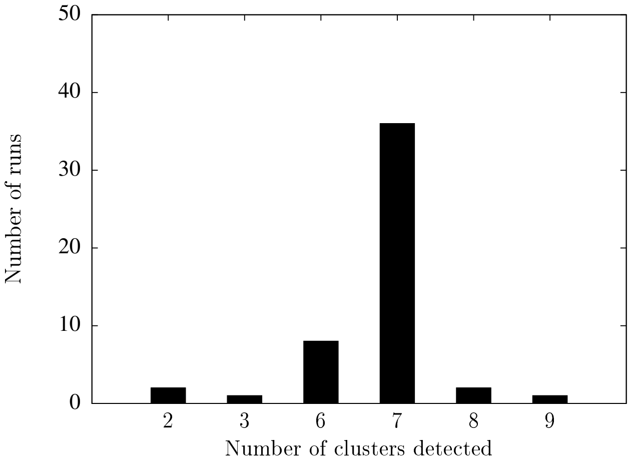

Figure 6.9 Histogram of the number of clusters detected in the hepta data set by LNN_{SDOT}

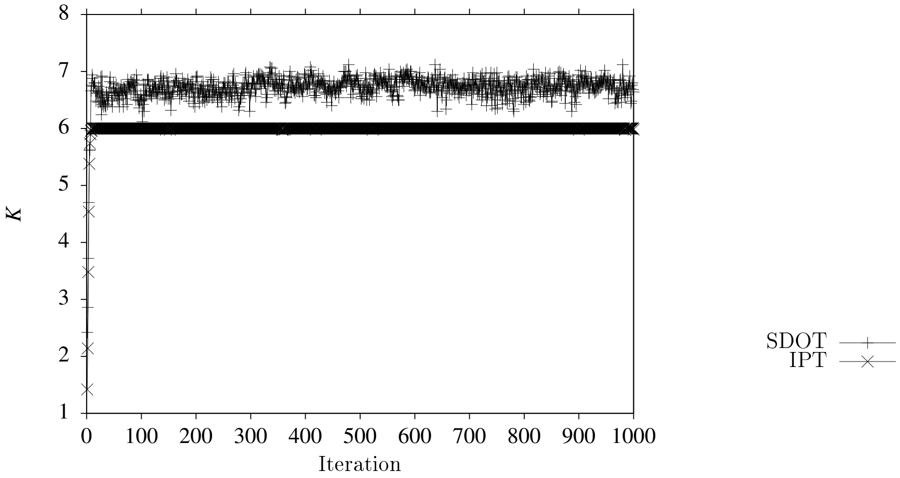

Figure 6.10 Convergence of LNNAIS using SDOT and IPT to optimal K for hepta data set

Table 6.4 Descriptive Statistics: Hepta

Algorithm	K	J_{intra}	J_{inter}	Q_{ratio}	Q_{DB}
KM$_{DB}$	7.00 (± 0.00)	0.993 (± 0.199)	4.041 (± 0.148)	1.112 (± 0.459)	0.870 (± 0.247)
KM$_{RT}$	4.00 (± 0.00)	1.680 (± 0.083)	3.902 (± 0.184)	0.630 (± 0.419)	1.006 (± 0.153)
LNN$_{DB}$	6.98 (± 0.14)	0.740 (± 0.122)	4.161 (± 0.097)	0.371 (± 0.259)	0.494 (± 0.219)
LNN$_{RT}$	5.98 (± 0.14)	1.019 (± 0.052)	4.307 (± 0.146)	0.316 (± 0.059)	0.661 (± 0.049)
LNN$_{SDOT}$	6.64 (± 1.21)	0.830 (± 0.397)	4.120 (± 0.231)	1.015 (± 4.978)	0.541 (± 0.365)

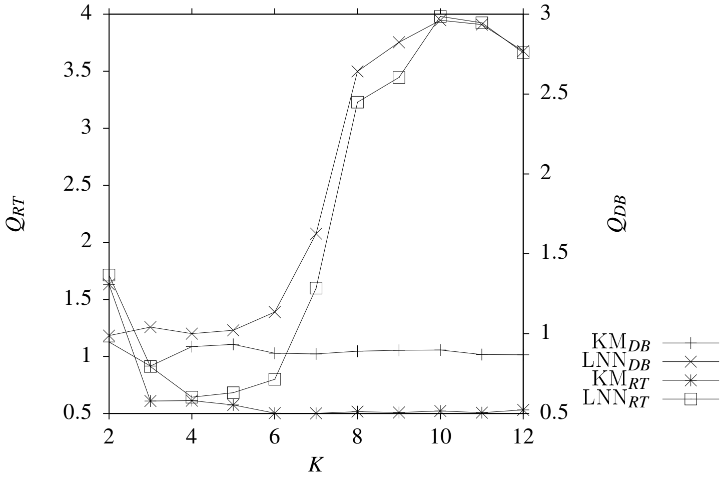

Figure 6.11 Optimal number of clusters obtained by K-means and LNNAIS for the engytime data set

has high frequencies for $2 \leq K \leq 4$ which is within the optimal range of K (refer to figure 6.12 for frequency distribution). Figure 6.13 illustrates that IPT obtains $K = 4$ for all iterations and SDOT oscillates around an average K of 4.4 over time for the engytime data set. There is no statistically significant difference between the clustering quality of any of the models (refer to table 6.12). Therefore, all models tend to deliver clusters with similar quality. LNN_{SDOT} has the advantage of dynamically determining the number of clusters in the engytime data set with similar clustering quality as the other models.

6.4.5 Chainlink data set

The optimal range of K for the chainlink data set is $[8,12]$ (as illustrated in figure 6.14). Figure 6.15 illustrates that LNN_{SDOT} has high frequencies for $K = 2$ and $4 \leq K \leq 7$ which are not within the optimal range of K. However, the figure also shows that there are cases where LNN_{SDOT} determined the number of clusters within the optimal range of K at lower frequencies. Note that the similarity between the range of determined clusters in figure 6.15 and the range of K for the iterative and multiple execution approaches in figure 6.14 is a coincidence. Figure 6.16 illustrates that IPT obtains $K = 8$ for all iterations and SDOT oscillates around an average K of 6.5 between $K = 5.5$ and $K = 8$ over time for the chainlink data set. The average number of clusters determined by LNN_{SDOT} for the chainlink data set is $K = 5.76$ (refer to table 6.6). Table 6.6 shows the results obtained by the different models to determine the optimal number of clusters in the chainlink data set.

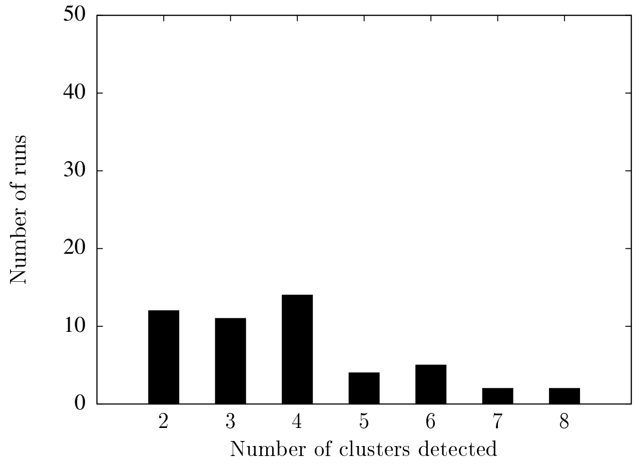

Figure 6.12 Histogram of the number of clusters detected in the engytime data set by LNN$_{SDOT}$

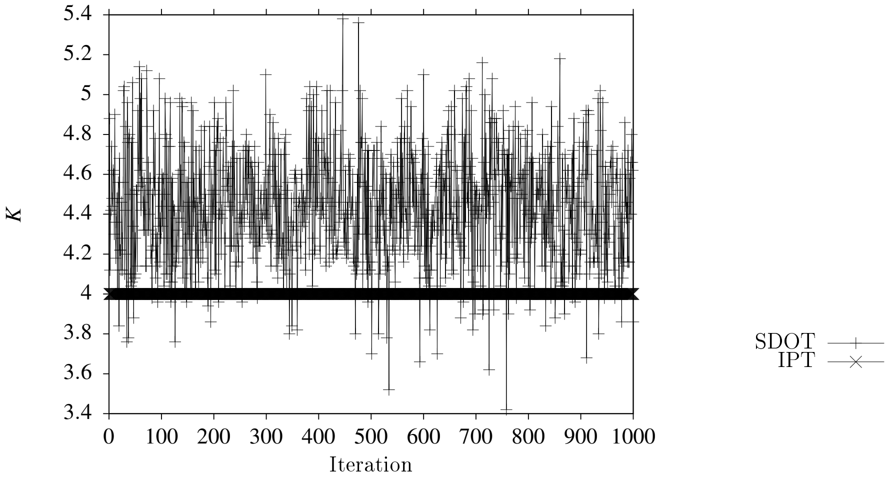

Figure 6.13 Convergence of LNNAIS using SDOT and IPT to optimal K for engytime data set

Table 6.5 Descriptive Statistics: Engytime

Algorithm	K	J_{intra}	J_{inter}	Q_{ratio}	Q_{DB}
KM_{DB}	3.00 (± 0.00)	1.165 (± 0.000)	3.184 (± 0.000)	0.396 (± 0.000)	0.797 (± 0.000)
KM_{RT}	7.00 (± 0.00)	0.805 (± 0.004)	3.188 (± 0.109)	0.502 (± 0.021)	0.873 (± 0.017)
LNN_{DB}	2.00 (± 0.00)	1.833 (± 0.213)	4.133 (± 1.032)	0.465 (± 0.107)	0.910 (± 0.194)
LNN_{RT}	4.00 (± 0.00)	1.284 (± 0.113)	4.020 (± 0.712)	0.616 (± 0.226)	1.000 (± 0.258)
LNN_{SDOT}	3.86 (± 1.62)	1.381 (± 0.304)	3.978 (± 0.808)	0.582 (± 0.217)	0.992 (± 0.287)

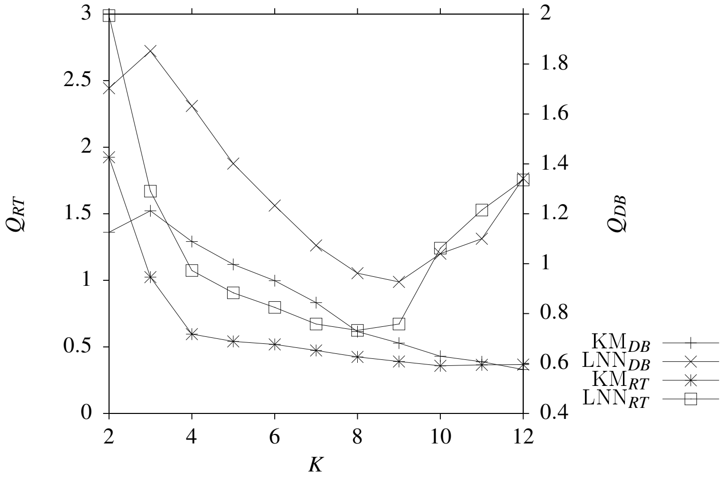

Figure 6.14 Optimal number of clusters obtained by K-means and LNNAIS for the chainlink data set

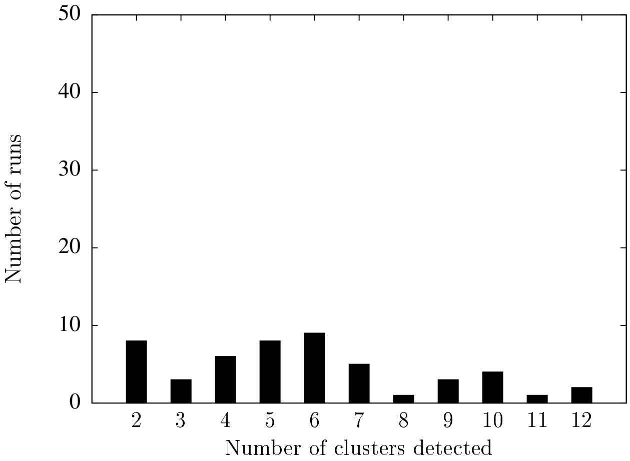

Figure 6.15 Histogram of the number of clusters detected in the chainlink data set by LNN$_{SDOT}$

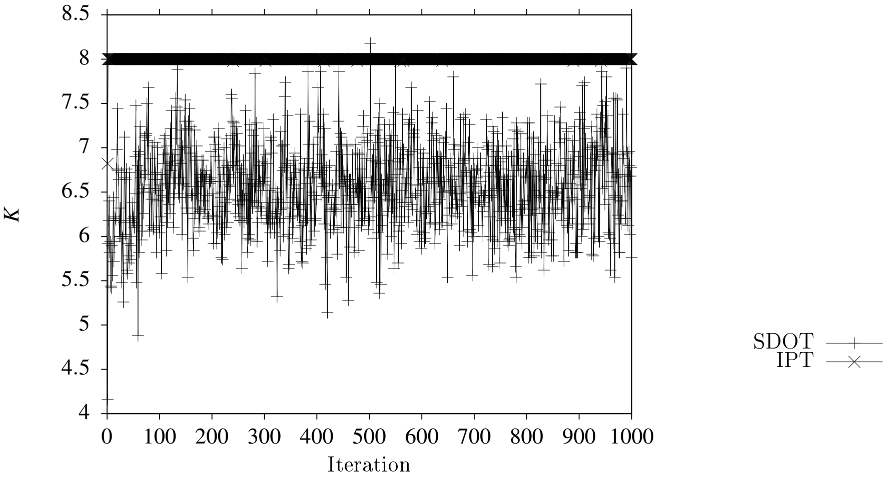

Figure 6.16 Convergence of LNNAIS using SDOT and IPT to optimal K for chainlink data set

Table 6.6 Descriptive Statistics: Chainlink

Algorithm	K	J_{intra}	J_{inter}	Q_{ratio}	Q_{DB}
KM$_{DB}$	12.00	0.262	1.500	0.367	0.576
	(± 0.00)	(± 0.009)	(± 0.025)	(± 0.063)	(± 0.017)
KM$_{RT}$	10.00	0.308	1.509	0.358	0.629
	(± 0.00)	(± 0.007)	(± 0.031)	(± 0.028)	(± 0.030)
LNN$_{DB}$	9.00	0.384	1.475	0.629	0.906
	(± 0.00)	(± 0.018)	(± 0.068)	(± 0.210)	(± 0.144)
LNN$_{RT}$	8.00	0.427	1.464	0.624	0.962
	(± 0.00)	(± 0.021)	(± 0.057)	(± 0.302)	(± 0.190)
LNN$_{SDOT}$	5.76	0.588	1.402	0.770	1.283
	(± 2.76)	(± 0.184)	(± 0.235)	(± 0.400)	(± 0.666)

Referring to table 6.12, the statistical hypothesis test rejects H_0 that the Q_{ratio} means are the same between KM$_{RT}$ and LNN$_{SDOT}$ ($z = 8.483$, $p < 0.001$). There is thus a statistical significant difference between the clustering quality of KM$_{RT}$ and LNN$_{SDOT}$. KM$_{RT}$ tends to find clusters in the chainlink data set with a higher quality than LNN$_{SDOT}$. There is also a statistical significant difference between the Q_{ratio} means of LNN$_{RT}$ and LNN$_{SDOT}$ ($z = 2.547$, $p = 0.011$). LNN$_{RT}$ tends to find clusters in the chainlink data set with a higher quality than LNN$_{SDOT}$.

6.4.6 Target data set

The average number of clusters determined by LNN$_{SDOT}$ for the target data set is $K = 4.04$ which is close to the optimal range of K (as illustrated in figure 6.17, $5 \leq K \leq 8$). The frequency distribution of the number of clusters determined by LNN$_{SDOT}$ for the target data set is illustrated in figure 6.18. LNN$_{SDOT}$ has high frequencies for $K \leq 5$. Figure 6.19 illustrates for the target data set the number of clusters respectively determined by the SDOT and IPT techniques over time. IPT obtains $K = 6$ for the majority of the iterations. The value of K for SDOT oscillates between $K = 3$ and $K = 5.5$ around an average K of 4.2 for the remaining iterations. Table 6.7 shows the results obtained by the different models to determine the optimal number of clusters in the target data set.

The statistical hypothesis test rejects H_0 that the Q_{ratio} means are the same between KM$_{RT}$ and LNN$_{SDOT}$ ($z = 7.835$, $p < 0.001$). There is thus a statistical significant difference between the clustering quality of KM$_{RT}$ and LNN$_{SDOT}$ and KM$_{RT}$ tends to find clusters in the target data

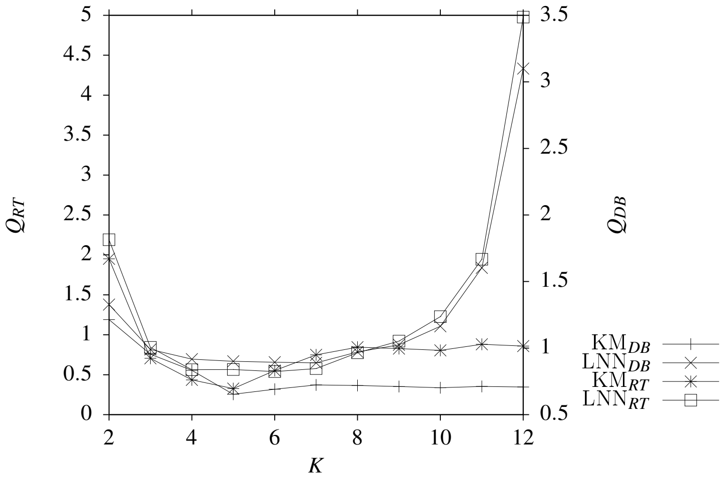

Figure 6.17 Optimal number of clusters obtained by K-means and LNNAIS for the target data set

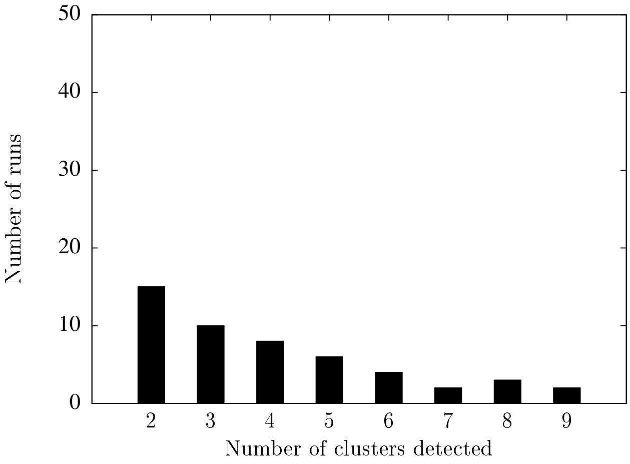

Figure 6.18 Histogram of the number of clusters detected in the target data set by LNN_{SDOT}

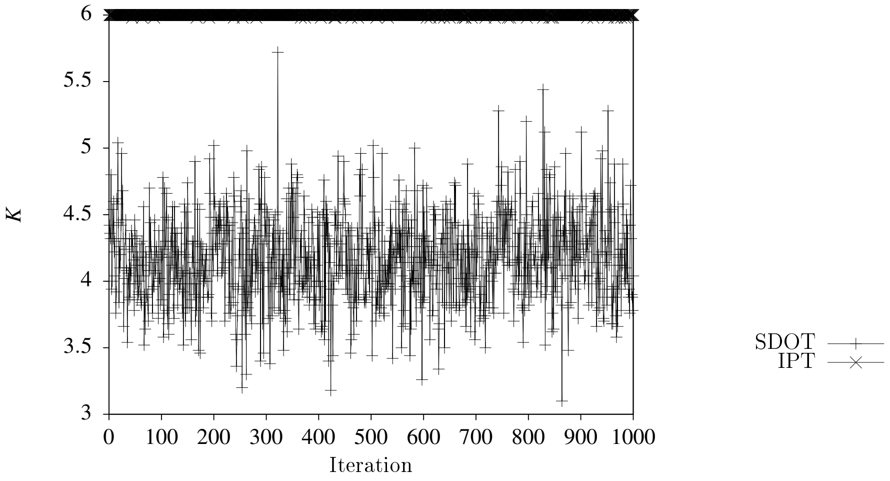

Figure 6.19 Convergence of LNNAIS using SDOT and IPT to optimal K for target data set

Table 6.7 Descriptive Statistics: Target

Algorithm	K	J_{intra}	J_{inter}	Q_{ratio}	Q_{DB}
KM$_{DB}$	5.00 (± 0.00)	0.533 (± 0.012)	2.313 (± 0.102)	0.326 (± 0.013)	0.653 (± 0.014)
KM$_{RT}$	5.00 (± 0.00)	0.533 (± 0.012)	2.313 (± 0.102)	0.326 (± 0.013)	0.653 (± 0.014)
LNN$_{DB}$	7.98 (± 0.14)	0.538 (± 0.075)	3.076 (± 0.343)	0.569 (± 0.477)	0.836 (± 0.284)
LNN$_{RT}$	6.00 (± 0.00)	0.661 (± 0.117)	2.806 (± 0.417)	0.539 (± 0.178)	0.894 (± 0.225)
LNN$_{SDOT}$	4.04 (± 2.04)	0.878 (± 0.208)	2.841 (± 0.751)	0.577 (± 0.438)	1.024 (± 0.860)

Table 6.8 Descriptive Statistics: Ionosphere

Algorithm	K	J_{intra}	J_{inter}	Q_{ratio}	Q_{DB}
KM$_{DB}$	2.00 (± 0.00)	2.289 (± 0.098)	3.156 (± 0.413)	0.730 (± 0.039)	1.484 (± 0.153)
KM$_{RT}$	4.00 (± 0.00)	2.085 (± 0.065)	3.438 (± 0.481)	0.877 (± 0.164)	1.776 (± 0.283)
LNN$_{DB}$	2.00 (± 0.00)	2.888 (± 0.278)	4.083 (± 0.642)	0.720 (± 0.100)	1.437 (± 0.257)
LNN$_{RT}$	5.00 (± 0.00)	2.473 (± 0.272)	4.277 (± 0.517)	0.911 (± 0.180)	1.755 (± 0.258)
LNN$_{SDOT}$	8.28 (± 2.12)	2.251 (± 0.322)	5.012 (± 0.424)	2.791 (± 6.519)	1.956 (± 1.737)

set with a higher quality than LNN$_{SDOT}$. There is however no statistical significant difference between the Q_{ratio} means of LNN$_{RT}$ and LNN$_{SDOT}$ (statistical hypothesis test accepts H_0, refer to table 6.12).

6.4.7 Ionosphere data set

Table 6.8 shows the results obtained by the different models to determine the optimal number of clusters in the ionosphere data set. Figure 6.20 illustrates that the optimal range of K for the ionosphere data set is $2 \leq K \leq 5$ (also shown in table 6.8). LNN$_{SDOT}$ determined the average number of clusters in the ionosphere data set as $K = 8.28$. The frequency distribution of the number of clusters determined by LNN$_{SDOT}$ for the ionosphere data set illustrates that LNN$_{SDOT}$ has high frequencies for $8 \leq K \leq 11$ which is not within the optimal range of K (refer to figure 6.21 for frequency distribution). Figure 6.22 illustrates for the ionosphere data set the number of clusters respectively determined by the SDOT and IPT techniques over time. The value of K for IPT rapidly increases to 5 in the first few iterations and remains at 5 for the majority of the remaining iterations. The value of K for SDOT rapidly increases to 8 and oscillates between $K = 7$ and $K = 9$ around an average K of 8 for the remaining iterations. Even though there is a difference in the optimal range of K between the models, there is no statistically significant difference between the clustering qualities of any of the models (refer to table 6.12). Therefore, all models tend to deliver clusters with similar quality at different optimal number of clusters. LNN$_{SDOT}$ has the advantage of dynamically determining the number of clusters in the ionosphere data set with similar clustering quality as the other models.

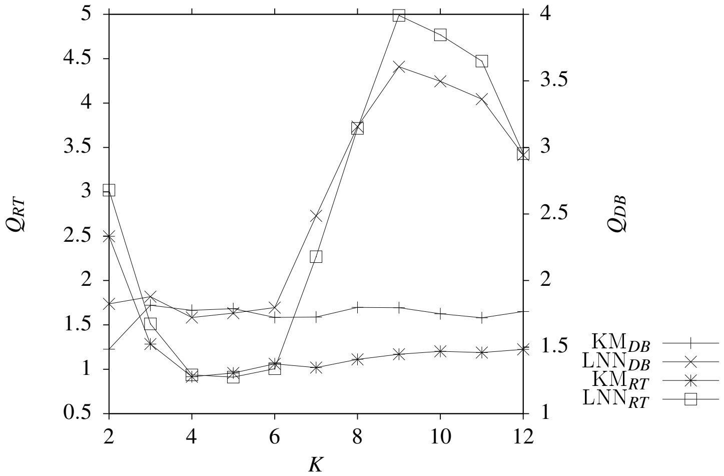

Figure 6.20 Optimal number of clusters obtained by K-means and LNNAIS for the ionosphere data set

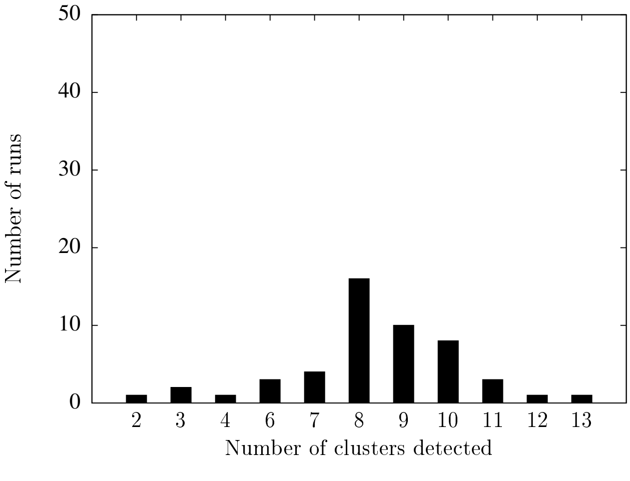

Figure 6.21 Histogram of the number of clusters detected in the ionosphere data set by LNN$_{SDOT}$

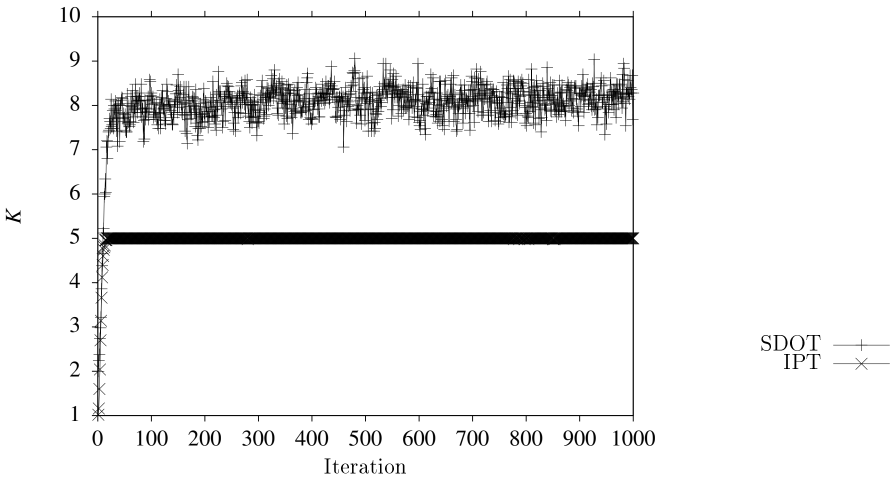

Figure 6.22 Convergence of LNNAIS using SDOT and IPT to optimal K for ionosphere data set

6.4.8 Glass data set

Figure 6.23 shows that the optimal number of clusters in the glass data set is obtained by KM$_{DB}$ and LNN$_{DB}$ at $K = 2$ and by KM$_{RT}$ and LNN$_{RT}$ at $K = 4$. Therefore the optimal range of K as determined by the different models for the glass data set is $[2,4]$. Figure 6.25 illustrates that the value of K for IPT rapidly increases to $K = 4$ and SDOT oscillates around an average K of 3.6 in range $[3,4.5]$ over time for the glass data set. Table 6.9 shows the results obtained by the different models to determine the number of clusters in the glass data set. The average number of clusters determined by LNN$_{SDOT}$ is $K = 3.34$ which falls within the optimal range of K.
A histogram of the frequency distribution of the number of clusters determined by LNN$_{SDOT}$ for the glass data set is illustrated in figure 6.24. LNN$_{SDOT}$ has high frequencies for $K \leq 5$. Referring to table 6.12, the Mann-Whitney U statistical hypothesis test rejects H_0 that the Q_{ratio} means are the same between KM$_{RT}$ and LNN$_{SDOT}$ ($z = 3.364$, $p < 0.001$) and between LNN$_{RT}$ and LNN$_{SDOT}$ ($z = 1.996$, $p = 0.046$). LNN$_{SDOT}$ tends to find clusters in the glass data set with a higher quality than KM$_{RT}$ and LNN$_{RT}$.

6.4.9 Image Segmentation data set

Table 6.10 shows the results obtained by the different models to determine the optimal number of clusters in the image segmentation data set. Figure 6.26 shows that the optimal number of clusters in the image data set is obtained by KM$_{DB}$ and LNN$_{DB}$ at $K = 2$, by KM$_{RT}$ at $K = 9$ and LNN$_{RT}$ at $K = 3$. The average number of clusters determined by LNN$_{SDOT}$ is $K = 3.28$

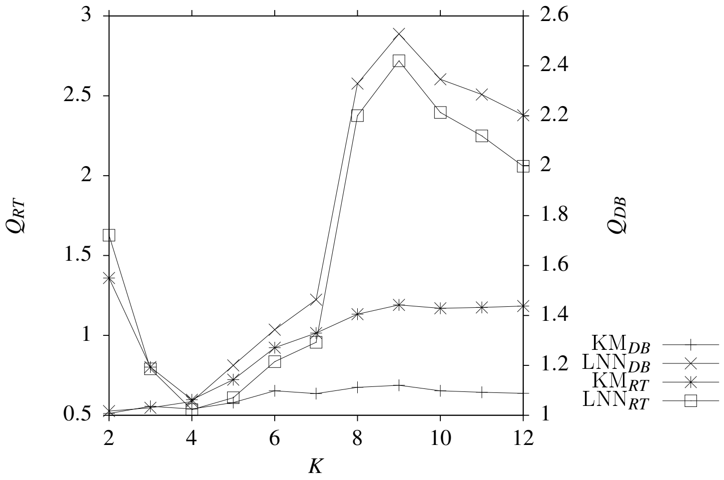

Figure 6.23 Optimal number of clusters obtained by K-means and LNNAIS for the glass data set

Table 6.9 Descriptive Statistics: Glass

Algorithm	K	J_{intra}	J_{inter}	Q_{ratio}	Q_{DB}
KM_{DB}	2.00 (± 0.00)	1.531 (± 0.100)	3.879 (± 0.546)	0.397 (± 0.019)	1.007 (± 0.116)
KM_{RT}	4.00 (± 0.00)	1.212 (± 0.056)	4.263 (± 0.627)	0.572 (± 0.152)	1.025 (± 0.149)
LNN_{DB}	2.00 (± 0.00)	2.354 (± 0.484)	5.792 (± 1.379)	0.427 (± 0.121)	0.892 (± 0.236)
LNN_{RT}	4.00 (± 0.00)	1.575 (± 0.208)	5.197 (± 0.769)	0.512 (± 0.161)	1.055 (± 0.266)
LNN_{SDOT}	3.34 (± 1.56)	2.003 (± 0.518)	5.998 (± 0.929)	0.493 (± 0.310)	0.875 (± 0.291)

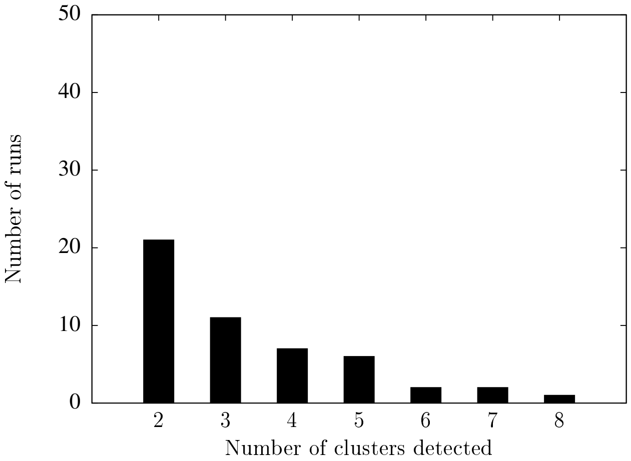

Figure 6.24 Histogram of the number of clusters detected in the glass data set by LNN$_{SDOT}$

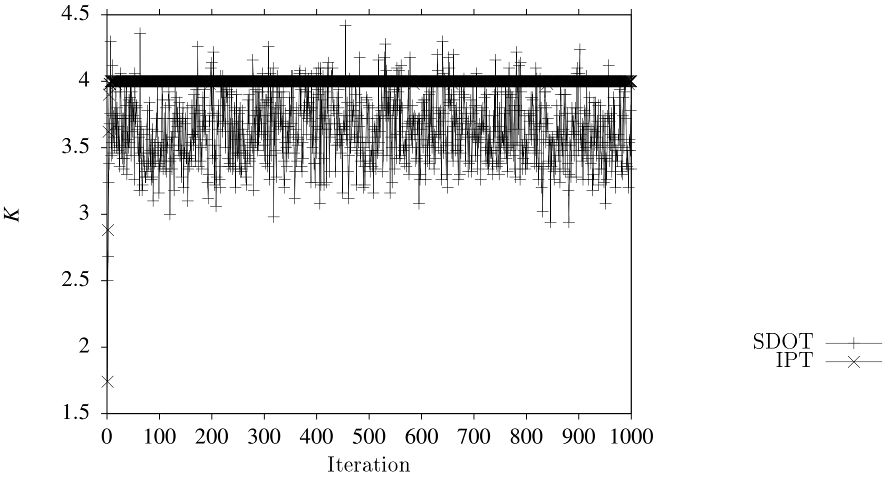

Figure 6.25 Convergence of LNNAIS using SDOT and IPT to optimal K for glass data set

Table 6.10 Descriptive Statistics: Image Segmentation

Algorithm	K	J_{intra}	J_{inter}	Q_{ratio}	Q_{DB}
KM$_{DB}$	2.00 (± 0.00)	101.487 (± 3.313)	238.922 (± 83.995)	0.439 (± 0.041)	0.861 (± 0.024)
KM$_{RT}$	9.00 (± 0.00)	58.442 (± 0.675)	322.656 (± 10.779)	0.688 (± 0.083)	1.021 (± 0.035)
LNN$_{DB}$	2.00 (± 0.00)	168.497 (± 33.481)	1148.026 (± 253.897)	0.155 (± 0.047)	0.551 (± 0.199)
LNN$_{RT}$	3.00 (± 0.00)	137.827 (± 15.718)	881.290 (± 141.104)	0.316 (± 0.253)	1.000 (± 0.602)
LNN$_{SDOT}$	3.28 (± 1.27)	142.847 (± 21.735)	975.017 (± 215.461)	43.919 (± 291.181)	88.577 (± 613.169)

which falls within the optimal range of K. Figure 6.28 illustrates that IPT obtains $K = 3$ for all iterations and SDOT oscillates around an average K of 3.2 in range $[2.6, 3.7]$ over time for the image data set. The frequency distribution of the number of clusters determined by LNN$_{SDOT}$ for the image segmentation data set is illustrated in figure 6.27. LNN$_{SDOT}$ has high frequencies for $K \leq 5$. Referring to table 6.12, the Mann-Whitney U statistical hypothesis test rejects H_0 that the Q_{ratio} means are the same between KM$_{RT}$ and LNN$_{SDOT}$ ($z = 6.89$, $p < 0.001$) and between LNN$_{RT}$ and LNN$_{SDOT}$ ($z = 2.337$, $p = 0.019$). LNN$_{SDOT}$ tends to find clusters in the image segmentation data set with a higher quality than KM$_{RT}$ and LNN$_{RT}$.

6.4.10 Spambase data set

The average number of clusters determined by LNN$_{SDOT}$ for the spambase data set is $K = 2.4$ which is within the optimal range of K (as illustrated in figure 6.29, $2 \leq K \leq 4$). In figure 6.29, note that $Q_{RT} < 0$ for LNN$_{RT}$ where $K \geq 10$. Q_{RT} values less than zero indicates that LNN$_{RT}$ was unable to cluster the data set into the corresponding K clusters. Since $\mathcal{B}_{max} = 10$ for data set spambase (refer to table 6.1), the number of clusters $K \geq 10$ is more than the number of available ALCs in the population. The frequency distribution of the number of clusters determined by LNN$_{SDOT}$ for the spambase data set is illustrated in figure 6.30. LNN$_{SDOT}$ has high frequencies for $K \leq 3$. Figure 6.31 illustrates that IPT obtains $K = 2$ for all iterations and SDOT oscillates around an average K of 2.45 in range $[2.2, 2.7]$ over time for the spambase data set. Table 6.11 shows the results obtained by the different models to determine the optimal number of clusters in the spambase data set. The statistical hypothesis test rejects H_0 that the Q_{ratio} means are the

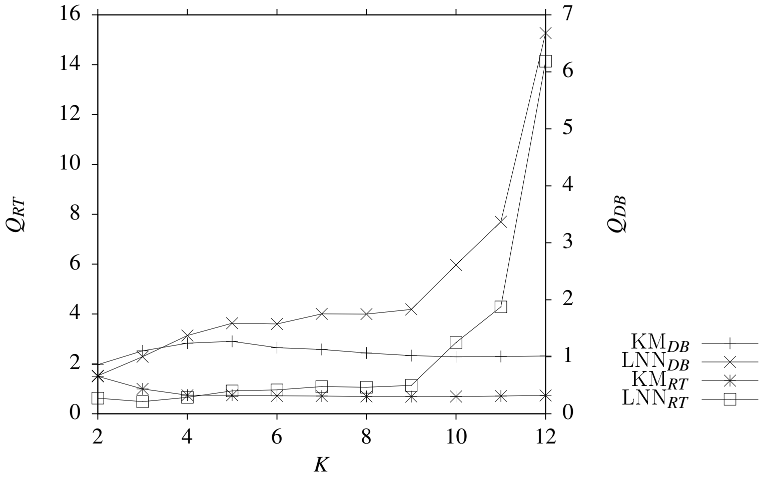

Figure 6.26 Optimal number of clusters obtained by K-means and LNNAIS for the image segmentation data set

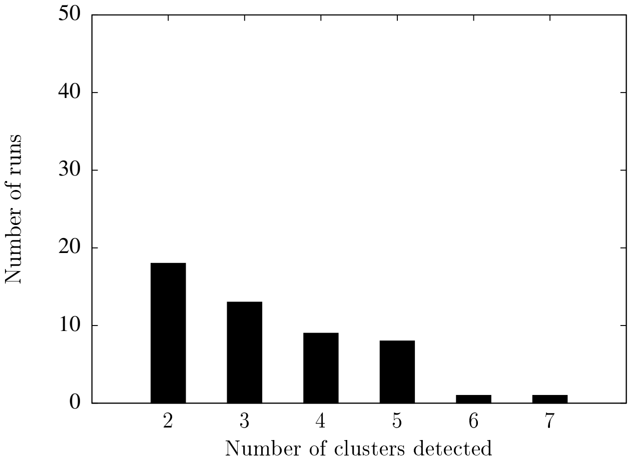

Figure 6.27 Histogram of the number of clusters detected in the image segmentation data set by LNN_{SDOT}

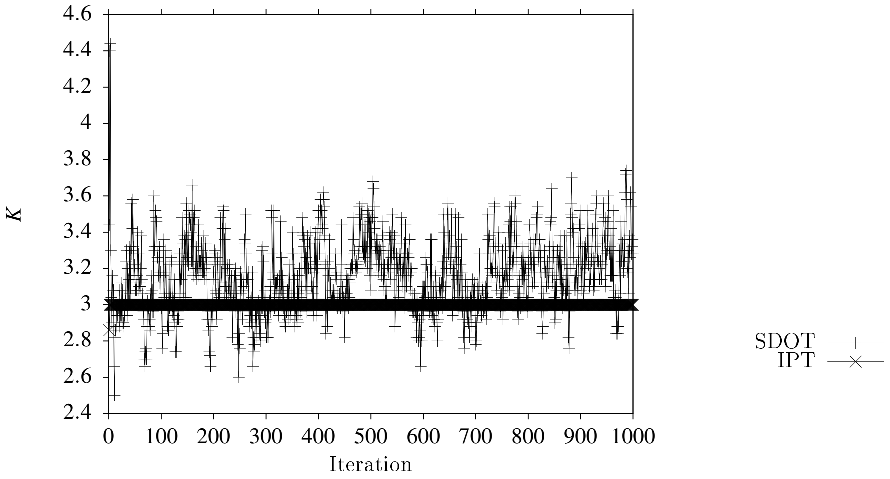

Figure 6.28 Convergence of LNNAIS using SDOT and IPT to optimal K for image data set

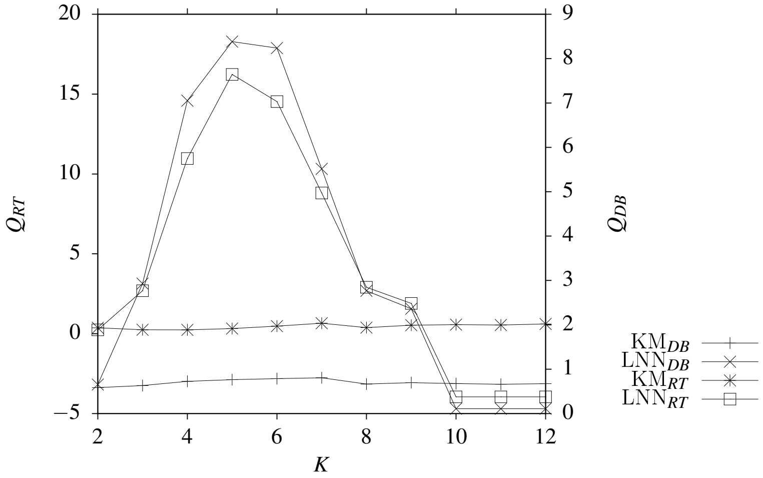

Figure 6.29 Optimal number of clusters obtained by K-means and LNNAIS for the spambase data set

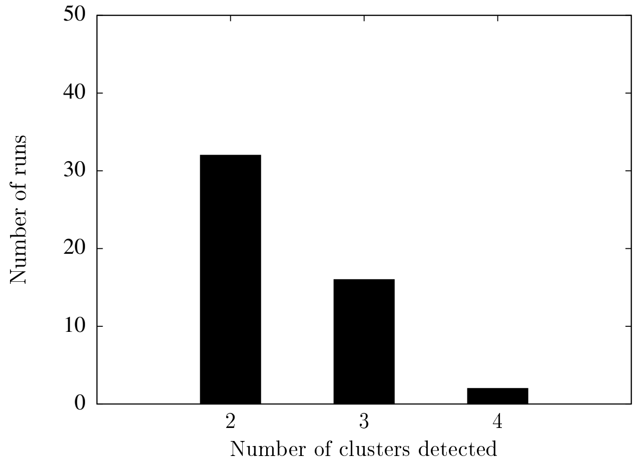

Figure 6.30 Histogram of the number of clusters detected in the spambase data set by LNN$_{SDOT}$

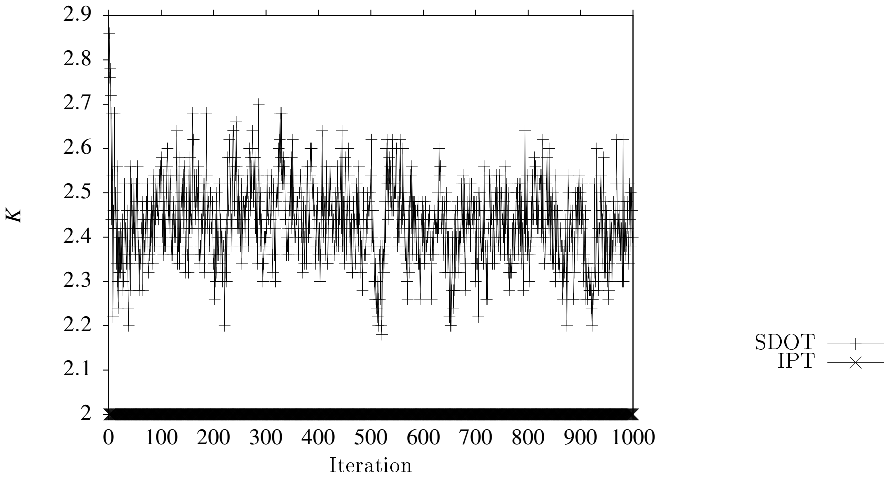

Figure 6.31 Convergence of LNNAIS using SDOT and IPT to optimal K for spambase data set

Table 6.11 Descriptive Statistics: Spambase

Algorithm	K	J_{intra}	J_{inter}	Q_{ratio}	Q_{DB}
KM$_{DB}$	2.00 (± 0.00)	216.058 (± 0.000)	2003.263 (± 0.000)	0.108 (± 0.000)	0.586 (± 0.000)
KM$_{RT}$	4.00 (± 0.00)	129.353 (± 0.000)	2165.832 (± 0.000)	0.229 (± 0.000)	0.727 (± 0.000)
LNN$_{DB}$	2.00 (± 0.00)	771.637 (± 317.716)	8288.589 (± 2462.940)	0.095 (± 0.031)	0.546 (± 0.077)
LNN$_{RT}$	2.00 (± 0.00)	475.834 (± 282.100)	7639.878 (± 2648.505)	0.071 (± 0.053)	0.655 (± 0.171)
LNN$_{SDOT}$	2.40 (± 0.57)	651.896 (± 382.136)	10416.929 (± 2798.913)	0.076 (± 0.042)	0.548 (± 0.222)

same between KM$_{RT}$ and LNN$_{SDOT}$ ($z = 8.269$, $p < 0.001$). There is thus a statistical significant difference between the clustering quality of KM$_{RT}$ and LNN$_{SDOT}$ and LNN$_{SDOT}$ tends to find clusters in the spambase data set with a higher quality than KM$_{RT}$. There is however no statistical significant difference between the Q_{ratio} means of LNN$_{RT}$ and LNN$_{SDOT}$ (statistical hypothesis test accepts H_0, refer to table 6.12).

For completeness, table 6.12 also shows whether there is a statistical significant difference between the clustering quality of KM$_{RT}$ and LNN$_{RT}$ for all the data sets. Referring to table 6.12, LNN$_{SDOT}$ and LNN$_{RT}$ tend to deliver clusters with a similar quality as KM$_{RT}$ for two of the data sets (engytime and ionosphere). Out of the remaining eight data sets, both LNN$_{SDOT}$ and LNN$_{RT}$ deliver clusters of a higher quality than KM$_{RT}$ for five of the data sets. Comparing LNN$_{SDOT}$ with LNN$_{RT}$ for five of the data sets (two-spiral, engytime, target, ionosphere and spambase) LNN$_{SDOT}$ tends to deliver clusters with a similar quality as LNN$_{RT}$. Out of the remaining five data sets, LNN$_{SDOT}$ delivers clusters of a higher quality than LNN$_{RT}$ for four of the data sets. In general, LNN$_{SDOT}$ tends to deliver clusters of similar or higher quality for all data sets, followed by LNN$_{RT}$ and KM$_{RT}$.

Table 6.12 Statistical Hypothesis Testing between All Models for all data sets based on Q_{ratio} as performance criteria ($\alpha = 0.05$; with continuity correction; unpaired; non-directional)

Data set	Model A	Model B	z of A	z of B	p	Outcome	Lowest z-score
iris	LNN$_{SDOT}$	KM$_{RT}$	-7.58	7.58	< 0.001	Reject H_0	LNN$_{SDOT}$
	LNN$_{RT}$	KM$_{RT}$	-3.209	3.209	0.001	Reject H_0	LNN$_{RT}$
	LNN$_{SDOT}$	LNN$_{RT}$	-6.69	6.69	< 0.001	Reject H_0	LNN$_{SDOT}$
two-spiral	LNN$_{SDOT}$	KM$_{RT}$	8.328	-8.328	< 0.001	Reject H_0	KM$_{RT}$
	LNN$_{RT}$	KM$_{RT}$	7.704	-7.704	< 0.001	Reject H_0	KM$_{RT}$
	LNN$_{SDOT}$	LNN$_{RT}$	-0.5	0.5	0.617	Accept H_0	LNN$_{SDOT}$
hepta	LNN$_{SDOT}$	KM$_{RT}$	-6.787	6.787	< 0.001	Reject H_0	LNN$_{SDOT}$
	LNN$_{RT}$	KM$_{RT}$	-8.145	8.145	< 0.001	Reject H_0	LNN$_{RT}$
	LNN$_{SDOT}$	LNN$_{RT}$	-4.391	4.391	< 0.001	Reject H_0	LNN$_{SDOT}$
engytime	LNN$_{SDOT}$	KM$_{RT}$	1.017	-1.017	0.309	Accept H_0	KM$_{RT}$
	LNN$_{RT}$	KM$_{RT}$	1.551	-1.551	0.121	Accept H_0	KM$_{RT}$
	LNN$_{SDOT}$	LNN$_{RT}$	-0.855	0.855	0.393	Accept H_0	LNN$_{SDOT}$
chainlink	LNN$_{SDOT}$	KM$_{RT}$	8.483	-8.483	< 0.001	Reject H_0	KM$_{RT}$
	LNN$_{RT}$	KM$_{RT}$	8.566	-8.566	< 0.001	Reject H_0	KM$_{RT}$
	LNN$_{SDOT}$	LNN$_{RT}$	2.547	-2.547	0.011	Reject H_0	LNN$_{RT}$
target	LNN$_{SDOT}$	KM$_{RT}$	7.835	-7.835	< 0.001	Reject H_0	KM$_{RT}$
	LNN$_{RT}$	KM$_{RT}$	8.145	-8.145	< 0.001	Reject H_0	KM$_{RT}$
	LNN$_{SDOT}$	LNN$_{RT}$	-0.221	0.221	0.825	Accept H_0	LNN$_{SDOT}$
ionosphere	LNN$_{SDOT}$	KM$_{RT}$	0.955	-0.955	0.340	Accept H_0	KM$_{RT}$
	LNN$_{RT}$	KM$_{RT}$	1.169	-1.169	0.243	Accept H_0	KM$_{RT}$
	LNN$_{SDOT}$	LNN$_{RT}$	0.283	-0.283	0.777	Accept H_0	LNN$_{RT}$
glass	LNN$_{SDOT}$	KM$_{RT}$	-3.364	3.364	< 0.001	Reject H_0	LNN$_{SDOT}$
	LNN$_{RT}$	KM$_{RT}$	-1.965	1.965	0.049	Reject H_0	LNN$_{RT}$
	LNN$_{SDOT}$	LNN$_{RT}$	-1.996	1.996	0.046	Reject H_0	LNN$_{SDOT}$
image segmentation	LNN$_{SDOT}$	KM$_{RT}$	-6.89	6.89	< 0.001	Reject H_0	LNN$_{SDOT}$
	LNN$_{RT}$	KM$_{RT}$	-7.18	7.18	< 0.001	Reject H_0	LNN$_{RT}$
	LNN$_{SDOT}$	LNN$_{RT}$	-2.337	2.337	0.019	Reject H_0	LNN$_{SDOT}$
spambase	LNN$_{SDOT}$	KM$_{RT}$	-8.269	8.269	< 0.001	Reject H_0	LNN$_{SDOT}$
	LNN$_{RT}$	KM$_{RT}$	-8.269	8.269	< 0.001	Reject H_0	LNN$_{RT}$
	LNN$_{SDOT}$	LNN$_{RT}$	1.275	-1.275	0.202	Accept H_0	LNN$_{RT}$

6.5 Influence of LNN$_{SDOT}$ Parameters

This section investigates the influence of the LNN$_{SDOT}$ parameters on the number of obtained clusters, K, in a data set. These parameters are the maximum population size, $\mathcal{B}_{max}$, the neighbourhood size, ρ, and the clonal level threshold, ε_{clone}. The influence of each parameter was evaluated for all the data sets listed in table 6.1 with the remaining parameters fixed at the values given in table 6.1.

Table 6.13: Effect of $\mathcal{B}_{max}$ on the number of detected clusters, K, by LNN$_{SDOT}$

Data set	$\mathcal{B}_{max}$	Optimal range	K
iris	10	$[2,4]$	2.48 ±0.608
	15		2.58 ±0.751
	20		2.84 ±0.857
	25		2.64 ±0.768
	30		2.88 ±1.070
	35		2.64 ±0.866
	40		2.88 ±0.952
two-spiral	10	$[3,12]$	3.46 ±1.445
	15		4.16 ±1.804
	20		4.06 ±1.891
	25		4.82 ±2.447
	30		4.56 ±2.410
	35		4.32 ±2.140
	40		5.40 ±3.521
hepta	10	$[4,7]$	4.28 ±1.470
	15		5.84 ±1.332
	20		6.18 ±1.571
	25		6.40 ±1.281
	30		6.60 ±1.149
	35		6.82 ±0.712
	40		6.64 ±1.213
engytime	10	$[2,7]$	3.32 ±1.009
	15		3.98 ±1.543
	20		3.86 ±1.625
	25		5.46 ±2.586
	30		5.20 ±3.013

Continued on next page

Data set	$\mathcal{B}_{max}$	Optimal range	K
	35		5.48 ±3.093
	40		6.34 ±4.043
chainlink	10	$[8,12]$	3.48 ±1.330
	15		4.64 ±2.278
	20		4.74 ±2.423
	25		6.54 ±3.145
	30		6.18 ±3.315
	35		5.78 ±3.472
	40		5.76 ±2.761
target	10	$[5,8]$	3.22 ±1.316
	15		4.16 ±1.901
	20		4.24 ±1.715
	25		4.08 ±2.505
	30		4.04 ±2.039
	35		3.96 ±2.433
	40		3.50 ±1.792
ionosphere	10	$[2,5]$	4.24 ±1.069
	15		6.40 ±1.510
	20		8.28 ±2.117
	25		10.16 ±2.716
	30		13.06 ±1.654
	35		15.72 ±2.764
	40		16.48 ±4.813
glass	10	$[2,4]$	3.22 ±1.238
	15		3.72 ±1.698
	20		3.34 ±1.557
	25		3.94 ±2.195
	30		3.68 ±2.083
	35		3.80 ±2.010
	40		3.94 ±2.378
	10		2.46 ±0.727

Continued on next page

Data set	$\mathcal{B}_{max}$	Optimal range	K
image	15	[2,9]	2.78 ±1.045
	20		3.08 ±1.197
	25		3.58 ±1.443
	30		3.28 ±1.266
	35		3.20 ±1.296
	40		3.52 ±2.823
spam	10	[2,4]	2.40 ±0.566
	15		2.82 ±1.260
	20		3.08 ±1.324
	25		2.90 ±0.900
	30		3.22 ±1.301
	35		3.24 ±1.305
	40		3.30 ±1.300

The influence of $\mathcal{B}_{max}$ was evaluated for all the data sets with the remaining parameters set to the values as listed in table 6.1. Table 6.13 summarises the results of the average detected K for each data set at different values of $\mathcal{B}_{max}$. There is a gradual to no increase in the number of obtained clusters, K, with an increase in $\mathcal{B}_{max}$ (as shown in table 6.13 for data sets iris, two-spiral, hepta, engytime, ionosphere, glass, image segmentation and spambase). There are also cases where K increases to a maximum and then starts to decrease with an increase in $\mathcal{B}_{max}$ (data sets chainlink and target). The effect of $\mathcal{B}_{max}$ on the number of obtained clusters for the ionosphere data set shows that $\mathcal{B}_{max} \geq 15$ tends to overfit the data since the number of obtained clusters is outside the optimal range. Therefore, the clustering performance of LNN$_{SDOT}$ with regards to K is sensitive to the value of $\mathcal{B}_{max}$.

Table 6.14: Effect of ε_{clone} on the number of detected clusters, K, by LNN$_{SDOT}$

Data set	ε_{clone}	Optimal range	K
	5		2.64 ±0.768

Continued on next page

Data set	ε_{clone}	Optimal range	K
iris	10	$[2,4]$	3.04 ±1.326
	15		3.08 ±1.383
	20		3.62 ±1.864
two-spiral	5	$[3,12]$	4.06 ±1.891
	10		4.92 ±2.415
	15		4.62 ±2.297
	20		5.14 ±2.136
hepta	5	$[4,7]$	6.64 ±1.213
	10		6.78 ±1.346
	15		6.66 ±1.365
	20		6.94 ±0.968
engytime	5	$[2,7]$	3.98 ±1.923
	10		3.86 ±1.625
	15		4.64 ±2.124
	20		4.40 ±1.811
chainlink	5	$[8,12]$	5.76 ±2.761
	10		6.76 ±3.456
	15		7.98 ±4.236
	20		7.68 ±4.420
target	5	$[5,8]$	4.04 ±2.039
	10		4.30 ±2.385
	15		4.24 ±2.526
	20		4.60 ±2.400
ionosphere	5	$[2,5]$	5.38 ±2.553
	10		7.50 ±1.652
	15		7.50 ±2.238
	20		8.28 ±2.117
glass	5	$[2,4]$	3.34 ±1.557
	10		4.32 ±1.794
	15		4.54 ±2.427
	20		4.56 ±2.080

Continued on next page

Data set	ε_{clone}	Optimal range	K
image	5	$[2,9]$	3.20 ±2.000
	10		3.08 ±1.573
	15		3.38 ±2.481
	20		3.28 ±1.266
spam	5	$[2,4]$	2.24 ±0.550
	10		2.32 ±0.546
	15		2.30 ±0.500
	20		2.40 ±0.566

The influence of ε_{clone} was evaluated for all the data sets with the remaining parameters set to the values as listed in table 6.1. Table 6.14 summarises the results of the average detected K for each data set at different values of ε_{clone}. There is a gradual or no increase in the number of obtained clusters, K, with an increase in ε_{clone} for all of the data sets (as shown in table 6.14). Therefore, the clustering performance of LNN$_{SDOT}$ with regards to K is sensitive to the value of ε_{clone}.

Table 6.15: Effect of ρ on the number of detected clusters, K, by LNN$_{SDOT}$

Data set	ρ	Optimal range	K
iris	3	$[2,4]$	2.64 ±0.768
	4		2.84 ±0.833
	5		2.58 ±0.724
two-spiral	3	$[3,12]$	4.06 ±1.891
	4		3.62 ±1.948
	5		4.46 ±3.517
hepta	3	$[4,7]$	6.64 ±1.213
	4		6.58 ±1.812
	5		5.80 ±2.307
engytime	3	$[2,7]$	3.86 ±1.625
	4		4.14 ±1.980

Continued on next page

Data set	ρ	Optimal range	K
	5		3.86 ±2.530
chainlink	3	$[8,12]$	5.76 ±2.761
	4		5.32 ±2.596
	5		5.10 ±3.775
target	3	$[5,8]$	4.04 ±2.039
	4		4.18 ±2.733
	5		4.20 ±2.828
ionosphere	3	$[2,5]$	8.28 ±2.117
	4		6.16 ±2.230
	5		5.12 ±1.935
glass	3	$[2,4]$	3.34 ±1.557
	4		3.76 ±2.006
	5		3.82 ±2.381
image	3	$[2,9]$	3.28 ±1.266
	4		3.48 ±1.910
	5		3.00 ±1.149
spam	3	$[2,4]$	2.60 ±0.800
	4		2.72 ±0.694
	5		2.40 ±0.566

The influence of ρ was evaluated for all the data sets with the remaining parameters set to the values as listed in table 6.1. Table 6.15 summarises the results of the average detected K for each data set at different values of ρ. There is generally no trend in the number of obtained clusters, K, with an increase in ρ except for the ionosphere data set where an increase in ρ decreases K (as shown in table 6.15). Therefore, the clustering performance of LNN$_{SDOT}$ with regards to K is generally insensitive to the value of ρ.

6.6 Conclusion

This chapter presented two techniques which can be used with LNNAIS to dynamically determine the number of clusters in a data set. These techniques are the iterative pruning technique (IPT) and the sequential deviation outlier technique (SDOT). Although both of these techniques are computationally less expensive than the multiple execution approaches, the IPT technique either needs a specified range for K or needs to iterate through all possible edges (to a maximum of $\mathcal{B}_{max}$) which makes the IPT technique parameter dependant in the former case and computationally slightly more expensive than SDOT in the latter. An advantage of IPT is that the technique can use any cluster validity index to determine the number of clusters. The SDOT technique neither uses a cluster validity index nor does it require any boundary constraints on K. SDOT is a non-parametric technique. This is an advantage, since it is not always feasible to visually inspect formed clusters, and a specified range for K might not contain the optimum number of clusters.

LNN$_{RT}$, LNN$_{DB}$ (both using IPT with Q_{RT} and Q_{DB}, respectively) and LNN$_{SDOT}$ (using SDOT) were applied on different data sets to determine the optimal number of clusters. These results were compared to the results obtained from K-means clustering which used the multiple execution approach to determine the optimal number of clusters in each data set. Based on the Q_{ratio} index, in general, LNN$_{SDOT}$ tends to deliver clusters of similar or higher quality for all data sets, followed by LNN$_{RT}$ and KM$_{RT}$. The influence of the different LNN$_{SDOT}$ parameters was also investigated.

Since the LNN$_{SDOT}$ model is computationally less expensive and is able to dynamically determine the number of clusters in a data set, the model can be seen as an enhancement to the LNNAIS model. Due to the possibility of the LNN$_{SDOT}$ model to dynamically determine the number of clusters, the model might indicate division or merging of clusters in a non-stationary environment. The next chapter defines and discusses different non-stationary environments and applies the proposed LNNAIS and LNN$_{SDOT}$ to the clustering of generated synthetic non-stationary data.