Using ART1 Neural Networks to Determine Clustering Tendency

Louis Massey

Royal Military College, Kingston, Ontario, Canada, K7K7B4

1. Introduction

Clustering is an unsupervised, data driven learning paradigm that aims at discovering natural groups in data [8, 9]. This type of learning has found many useful applications in domains with large amount of data where labeling of a training set for supervised learning is cost prohibitive or where autonomy is essential [1, 10, 11, 12]. However, clustering algorithms generally rely on some prior knowledge of the structure present in a data set. For instance, one needs to know whether or not clusters actually exist in data prior to applying a clustering procedure. Indeed, clustering applied to a data set with no naturally occurring clusters would merely *impose* meaningless structure. The procedure that consists in examining a data set to determine if structure is actually present and thus determine if clustering is a worthwhile operation is a poorly investigated problem known as *cluster tendency* determination [8].

Research in the area of cluster tendency has mainly focussed on the somewhat related problem of establishing the true number of clusters present in the data [6], often as part of *cluster validity*, the evaluation of clustering output quality [8]. Of course, should it be ascertained that the best clustering contains only one group, then null tendency must be concluded. The main problem with these approaches is that they either rely on yet other optimization procedures and similarity metrics (just as the clustering procedure itself), or depend on some parameter estimation. We show how to avoid these problems by using Adaptive Resonance Theory (ART) neural networks [3, 7] to determine clustering tendency of binary data. The binary version of ART (ART1) is used.

2. Adaptive Resonance Theory

ART neural networks are known for their ability to perform plastic yet stable online [14] clustering of dynamic data sets [4]. ART detects similarities among data objects, typically data points in an N-dimensional metric space. When novelty is detected, ART adaptively and autonomously creates a new category. Another ad-

vantageous and distinguishing feature of ART is its ability to discover concepts in data at various levels of abstraction [12, 16]. This is achieved with the *vigilance parameter* $\rho \in (0,1]$. First, a similarity measure S (Eq. 1) is computed to determine if an existing cluster prototype T_j appropriately represents the critical features of an input pattern X_k.

$$S = \|X_k \wedge T_j\| / \|X_k\| \tag{1}$$

Then, the *vigilance test* compares S with the vigilance parameter (Eq. 2):

$$S \geq \rho \tag{2}$$

Eq. 2 determines whether the current input pattern X_k will be recognized as a known concept or as a novel one. Indeed, if the vigilance test fails for all existing prototypes during the network search phase, X_k is deemed to be novel and new concept formation is triggered. At high vigilance ($\rho \to 1$), a large number of specific (low generality) clusters will be detected in the data. Conversely, at low vigilance ($\rho \to 0$), objects will be assigned to fewer, more general categories. Given a data set $X = \{X_k \mid k = 1, 2, \ldots, R\}$, one then intuitively expects a function relating the number of clusters M to the vigilance ρ. This function is expected to have a minimum value of M=1 for $\rho \to 0$ and a maximum of M=R for ρ=1.

3. Minimal and Maximal Vigilance

Of interest is S_{min}, the minimal non-zero value for S. The minimum non-zero value for the numerator of Eq. 1 is 1, that is one common bit between the prototype and the input data. The theoretical[1] maximal value for the denominator is N. Hence, we obtain:

$$S_{min} = 1/N. \tag{3}$$

Similarly, we develop a non-unit maximal value for S:

$$S_{max} = (N-1)/N. \tag{4}$$

Based on Eqs. 2, 3 and 4, one can establish the corresponding *minimal and maximal useful values* for vigilance. This is illustrated as follows. Lets suppose M_0 clusters are obtained with ρ_0. However, the application requires $M_1 < M_0$ clusters, so the input set is re-submitted to the ART1 neural network with $\rho_1 < \rho_0$. The expectation is that with a lower vigilance, fewer and more general groups will be formed: this is normal ART1 behavior. Under certain conditions, this will not occur. Indeed, if $\rho_1 < S_{min}$, then any further reduction of the vigilance will not result in less clusters[2]. The last vigilance value for which a reduction in the number of clusters was achieved is the minimal useful vigilance, ρ_{min}. A similar reasoning ap-

[1] In practice $\|X_k\|$ is expected to be a small fraction of N, so the actual S_{min} (Eq. 3) will always be larger or equal to the theorical value, but this is inconsequent for the use we make of it.

[2] Under certain circumstances, the number of clusters may decrease slightly [15].

plies for the maximal useful vigilance ρ_{max} in which case a maximum number of clusters $M \leq R$ is reached. Hence, in reality, the intuition that varying the vigilance within its whole range of permissible value can give rise to an arbitrary number of clusters $M \in [1,R]$ is incorrect.

Minimal and maximal useful vigilance may be problematic for applications that require concept granularities beyond or below possible limits. However, they are also useful structure detector. Indeed, the phenomenon just described can be interpreted as an inherent ability of ART1 to detect natural similarities in the data. For instance, *less clusters cannot be discovered in the data below minimal vigilance simply because the data is naturally not susceptible to assemble into less group*s.

4. Tendency Determination

By setting vigilance to a value below minimum useful vigilance, one obtains an indication of the natural tendency of the data to cluster. This is deduced from the number of clusters formed. The following situations can occur for $\rho<\rho_{min}$:

1. If M>1: it can be concluded that the data has a natural tendency to cluster.
2. If M=1: from the most general point of view[3], the data does not have a tendency to form clusters.
3. If $M \rightarrow R$: this means weak tendency, with the extreme case of M=R clusters (trivial clustering) corresponding to no tendency.

For situation 2, one must progressively increment the vigilance parameter from its minimal useful value to determine if eventually M>1 clusters form. Hence, one must consider the notion of *generality* when computing clustering tendency. This can be illustrated by considering fauna classification: at the highest level of the hierarchy (the more general cluster), all animals are clustered into category “animal” (i.e. M=1). However, this does not mean that there is no structure in the data. By lowering the generality, M>1 potentially useful clusters may be found (for example, corresponding to animal classes or families).

We have described a method to establish if natural groups occur in the data. The residual question is whether those groups are the result of mere coincidence. Indeed, it can easily be demonstrated [13] both analytically and empirically that clusters do occur in a random data set. Evidently, such clusters are meaningless and clustering tendency of such origin must be appropriately detected. We now show how maximal vigilance is used for that purpose.

Increasing vigilance means that the ART network will form more and smaller clusters since it is being more demanding about features matching, as per the vigilance test (Eq. 2). Random data should therefore have a natural predisposition to split into many small clusters more rapidly than data that contains actual structure because it is less likely to have the required number of bit matching to pass the more stringent vigilance test. Maximal vigilance for random data, which we

[3] Recall that ρ_{min} implies the more general clustering possible.

denote by ρ^{r}_{max}, will thus be smaller than for non-random. This idea can be used to establish non-random tendency as shown by the following inequality:

$$\rho^{r}_{max} < \rho_{max} \quad (5)$$

Maximal vigilance can be approached incrementally for a given data set and the relation between M and ρ plotted. Observing how fast M tends towards R compared to a baseline random data set allows for the detection of clustering tendency of a random origin. Such a graphic and qualitative approach may not be ideal, but it suits our current objective by giving an idea of whether or not clusters may be due to mere chance. Elements of a quantitative approach are given in [13].

5. Empirical Validation

Patterns are bit strings of length N. In the first experiment, we consider the case where tendency is determined by failing to reach M=1 at $\rho < \rho_{min}$. The data set with R=10 patterns and N=50 (Fig. 1a) is submitted to ART at $\rho < \rho_{min}$ ($\rho = 0.01$), then re-submitted at progressively incremented vigilance. The effect of minimal vigilance is visible in Fig. 1b. M=3 clusters are detected at below minimum vigilance, which allows one to conclude that there is clustering tendency for the data set. Visual inspection of figure 1a confirms this finding. Applying a clustering procedure to the data is therefore a meaningful operation.

In the second experiment, any data set in which not a single feature overlap exists can be used. This kind of data has no inherent clustering tendency due to an absolute lack of similarity between objects. By clustering this data at below minimum vigilance, we obtain M=R, hence confirming null tendency. Applying a clustering procedure to this data set would impose artificial structure and would therefore be a meaningless operation.

In the third experiment, we verify that the rate at which M grows when ρ is progressively incremented to ρ_{max} detects tendency caused by chance as per Eq. 5. Three random[4] data sets are processed with ART at vigilance varying from below ρ_{min} to 1. The number of clusters formed is averaged for each vigilance value. Other data sets will be compared with this baseline to determine if their clustering

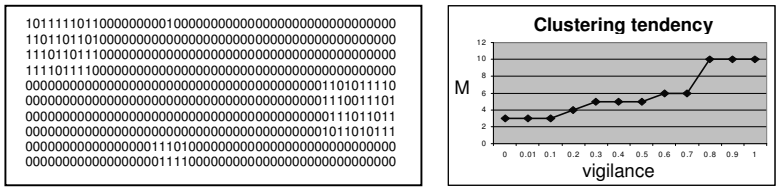

Fig. 1a. Dataset for experiment 1. **Fig. 1b.** Clustering tendency is established by observing that 3 clusters are formed at vigilance below minimum useful vigilance. The number of clusters formed at various levels of generality is also visible.

[4] The random bit patterns are actually pseudo-random data generated with java.util.Random.

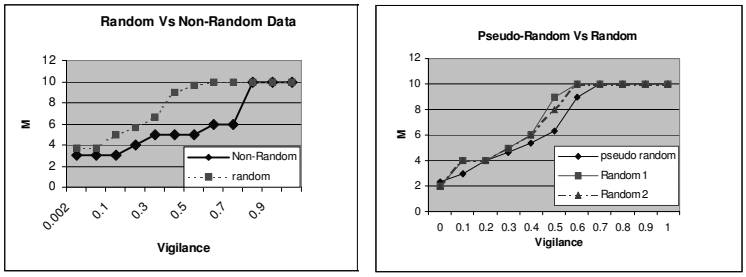

Fig. 2a. Baseline pseudo-random data reaches maximal vigilance faster than non-random data. This indicates that clustering tendency is not caused by chance clustering **Fig. 2b.** True random data reach maximal vigilance faster than the baseline, which is an indication that their clustering tendency is caused by mere chance.

tendency is caused by chance. The random data must possess the same characteristics (N, R and P(1), the bit set probability) as the dataset to be compared against. Here we used N=50, R=10 and P(1)=0.2. In Fig. 2a, the number of clusters is plotted for the random baseline and for the data used in experiment 1. It can be observed that ρ^{r}_{max} for the random baseline is smaller than ρ_{max} for the tested data. As per Eq. 5, this is an indication that cluster tendency is not caused by random structure in the data. In Fig. 2b, two other data sets are compared to the random baseline. For these data sets, $\rho^{r}_{max} > \rho_{max}$, which means that clustering tendency is caused by mere chance clustering. Indeed, these two data sets were obtained from known random sources (random 1 comes from radio atmospheric noise and was obtained at random.org. Random 2 comes from radioactive source decays and was obtained from HotBits (http://www.fourmilab.ch/hotbits/)). Other similar experiments have been conducted with several real-life or benchmark data sets. One of these experiments is documented in [12].

6. Conclusion and Future Work

We have shown how the vigilance parameter of a binary ART neural network can be used to determine the clustering tendency of a data set. The idea is based on the fact that at the highest level of generality, that is for vigilance set below its minimal theoretical level, ART should collapse all clusters into a single group. However, if the data possesses inherent structure, it will not. It was furthermore argued that clustering tendency can be achieved at various levels of generality. We also described the use of maximal vigilance to detect cluster tendency caused by chance. Hence, tendency is determined in two simple steps: first, verify that the data does not contain a trivial number of clusters (M=1 or M=R) at minimal vigilance; and second, verify that non-trivial clustering are not caused by chance by considering the rate at which maximum vigilance is reached compared to baseline random data. The method to determine clustering tendency as described in this paper is applicable to binary data inputs only; investigation of the non-binary ART

versions, such as ART2 [2] and fuzzyART [5] on real-valued continuous data would be an interesting area of future research.

References

[1] Bezdek JC (1993) Review of MRI Images Segmentation Techniques Using Pattern Recognition. Medical Physics vol 20 no 4 pp1033-1048

[2] Carpenter GA, Grossberg S (1987) Art 2: Self-organisation of stable category recognition codes for analog input patterns. Applied Optics Vol 26 pp 4919-4930

[3] Carpenter GA, Grossberg S (1987) Invariant pattern recognition and recall by an attentive self-organizing ART architecture in a nonstationary world. In: Proceedings of the IEEE First International Conference on Neural Networks pages II-737-745

[4] Carpenter GA, Grossberg S (1995) Adaptive Resonance Theory (ART). In: Handbook of Brain Theory and Neural Networks Ed: Arbib MA, MIT Press

[5] Carpenter GA, Grossberg S, Rosen DB (1991) Fuzzy art: Fast stable learning and categorization of analog patterns by an adaptive resonance system. Neural Networks Vol 4 pp 759-771

[6] Dubes RC (1987) How Many Clusters are Best? – An Experiment. Pattern Recognition Vol 20 No 6 pp 645-663

[7] Grossberg S (1976) Adaptive pattern classification and universal recording : I Parallel development and coding of neural feature detectors. Biological Cybernetics Vol 23 pp 121-134

[8] Jain AK, Murty MN, Flynn PJ (1999) Data Clustering: A Review. ACM Computing Surveys Vol 31 No 3

[9] Kaufman L, Rousseeuw PJ (1990) Finding Groups in Data: An Introduction to Cluster Analysis. Wiley-Interscience

[10] Kohonen T, Lagus K, Salojärvi J, Honkela J, Paatero V, Saarela A (2000) Self Organization of a Document Collection. IEEE Transactions On Neural Networks Vol 11 No 3

[11] Li C, Biswas G (1995) Knowledge-based scientific discovery in geological databases. In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining, Montreal, Canada, pp 204–209

[12] Massey L (2002) Structure Discovery in Text Collections. In: Proc of KES'2002 Sixth International Conference on Knowledge-Based Intelligent Information & Engineering Systems, Podere d'Ombriano, Crema, Italy, pp161-165

[13] Massey L (2002) Determination of Clustering Tendency With ART Neural Networks. In: Proceedings of 4th Intl. Conf. on Recent Advances in Soft Computing, Nottingham, U.K., 12 & 13 December 2002.

[14] Moore B (1988) ART and Pattern Clustering. Proceedings of the 1988 Connectionist Models Summer School pp174-183

[15] Sadananda R, Sudhakara Rao GRM. (1995) ART1: model algorithm characterization and alternative similarity metric for the novelty detector. In: Proceedings IEEE International Conference on Neural Networks Vol 5 pp2421 –2425

[16] Vlajic N, Card H-C (1998) Categorizing Web Pages using modified ART. In: Proceedings of IEEE 1998 Canadian Conference on Electrical and Computer Engineering Waterloo, Canada Vol1 pp313-316