Nuclear Potential Clustering As a New Tool to Detect Patterns in High Dimensional Datasets

V Tonkova1, D Paulus2, H Neeb1
1RheinAhrCampus Remagen, University of Applied Sciences Koblenz, Süddelee 2, 53424 Remagen, Germany
2University of Koblenz-Landau, Faculty of Computer Science, Universitätsstraße 1, 56070 Koblenz
E-mail: tonkova@rheinahrcampus.de

Abstract. We present a new approach for the clustering of high dimensional data without prior assumptions about the structure of the underlying distribution. The proposed algorithm is based on a concept adapted from nuclear physics. To partition the data, we model the dynamic behaviour of nucleons interacting in an N-dimensional space. An adaptive nuclear potential, comprised of a short-range attractive (strong interaction) and a long-range repulsive term (Coulomb force) is assigned to each data point. By modelling the dynamics, nucleons that are densely distributed in space fuse to build nuclei (clusters) whereas single point clusters repel each other. The formation of clusters is completed when the system reaches the state of minimal potential energy. The data are then grouped according to the particles’ final effective potential energy level. The performance of the algorithm is tested with several synthetic datasets showing that the proposed method can robustly identify clusters even when complex configurations are present. Furthermore, quantitative MRI data from 43 multiple sclerosis patients were analyzed, showing a reasonable splitting into subgroups according to the individual patients’ disease grade. The good performance of the algorithm on such highly correlated non-spherical datasets, which are typical for MRI derived image features, shows that Nuclear Potential Clustering is a valuable tool for automated data analysis, not only in the MRI domain.

1. Introduction
Cluster analysis is one of the standard techniques to analyze and blindly group data into different categories. However, exploring natural phenomena has often to do with high dimensional correlated non-spherical distributed features, which confine the application of standard clustering methods [1]. A current problem in the neuroscience, for example, is the extraction of image-based biomarkers for the diagnosis of Multiple Sclerosis (MS), a neurodegenerative disease leading to a progressive lifelong handicap. Here, a mismatch between conventional medical imaging and clinical manifestation of the disease is observed [2, 3]. Even though new quantitative magnetic resonance imaging techniques that offer the possibility of measuring multiple parameters simultaneously have emerged [2, 6], no sensitive biomarkers have yet been extracted, since the values measured in a multimodal MRI sequence are biophysically correlated.

In order to detect patterns in high dimensional datasets, we developed a new clustering algorithm, which does not rely on a priori assumptions about the distribution of the data, i.e., correlated data do not pose severe problems. The basic idea is to treat the points to be classified as nucleons in an N-
dimensional space and model their dynamic behaviour. A nuclear potential is associated to every ‘particle’, so that, as times goes on, points that are densely distributed in space fuse to form ‘nuclei’, i.e. clusters.

To validate the algorithm, correlated test datasets of different dimensionality were simulated and analyzed. The method was then applied on a dataset, consisting of image based features derived from spatially normalized $T_1^*, T_2^*$, and absolute water content maps of 43 MS patients. The results showed a very good performance of the Nuclear Potential Clustering on high dimensional correlated datasets.

2. Materials and Methods

2.1. Nuclear Potential Clustering

The Nuclear Potential Clustering method is based on the modelling of a physical system in which each data point of an N-dimensional dataset represents a single nucleon. As such it possesses: (1) a short range potential well, defined here by a Woods-Saxon approach

$$\phi^s = \frac{-V_0}{r^{-1} a + 1},$$  \hspace{1cm} (1)

that traps nearby nucleons (strong interaction) and (2) at separations larger than the width of the potential well ($r > R$) a positive Coulomb potential $\phi^C = r^{-1}$ (electromagnetic interaction, where the electric charge is normalized to $4\pi \varepsilon_0$) [5]. Here, $r$ defines the Euclidean distance between two points in feature space, $R$ is the radius at which the strong interaction is effective and $a$ sets the slope of the potential at the border between the two terms [7]. The interaction strength is given by the $|V_0|$. Prior to clustering, features are normalized to have zero mean and unit standard deviation in order to provide an equal weight to each dimension when calculating distances. The basic idea underlying our clustering approach is now the following: while the force between nearby particles is dominated by the strong interaction, resulting in a mutual attraction, particles at larger distances experience only the repelling electromagnetic force. Thus, points that are close enough fuse to larger nuclei (clusters), while single nucleons (outliers) are repelled when dynamically modelling the system. By testing the performance of the algorithm on artificial data sets with different number, shape, density and dimensionality of the clusters, the values characterizing the potential function were empirically set to: $V_0 = 3500$ [a.u.], $R = 10\%$ quantile of the initial distance distribution between all points, and $a = R/50$. The range of the strong interaction is thus adapted to the data set to be clustered.

Starting from the initial spatial distribution of the nucleons, we explore the time evolution of the system. The effective force acting on a single particle is given by the sum of two terms: (1) the negative gradient of the total potential function $\phi(x)$ resulting from the interaction with all other points $\vec{F}_n = -\nabla \phi(\vec{x})$, and (2) a friction term proportional to the particles velocity, $\vec{F}_f = -\eta \cdot \vec{v}$. Setting the damping constant to $\eta > 0$ suppresses the scattering of nearby particles if their kinetic energy is large enough to leave the strong potential well. Furthermore, it leads to a quicker transition to equilibrium state, making the algorithm significantly faster. $\eta$ is set to 200 [a.u.], whereby the algorithm is largely immune against variations of this parameter. Given the total force, the motion of a point $\vec{x}(t)$ is obtained by solving the Newtonian equation of motion, using a fourth-order Runge-Kutta method:

$$m \cdot \ddot{\vec{x}}(t) = \vec{F}(t) - \eta \cdot \dot{\vec{x}}(t), \text{ with } \dot{\vec{x}}(t) = \vec{v}(t) \text{ and } \ddot{\vec{x}}(t) = \dot{\vec{x}}(t).$$  \hspace{1cm} (2)

Here $m$ represents the mass of a particle which we set to 1. After every Runge-Kutta step, the total potential and kinetic energy of every particle are determined.

We have observed that points do not directly form stable nuclei but rather oscillate towards an equilibrium state. In order to choose an objective stopping condition for our algorithm, we determine the system temperature as a function of time, $T(t)$ (an example for a test data set is shown in Figure 1c)). The temperature $T$ is given by the average kinetic energy of all particles. The algorithm is stopped when $T$ reaches a value less than 1% of the absolute maximum reached. Furthermore, the repelling Coulomb force results in a splitting of larger clusters into sub-clusters at locations where the particle density decreases. In order to prevent from such an artificial splitting, we increase the value of $R$ for times later than the time point, where the first temperature maximum is reached.
\[ R(t) = R(t_0) \cdot (1 + a t^{-1}) \], \[ a_t = 4, 5, 6 \ldots \] (3)

The algorithm is repeated five times with different initial particle velocities \( \theta(t_0) \) in order to reduce the incidental assignment of noise points to one of the clusters formed. During the first run, particles start at rest. For the next repetitions \( \theta(t_0) \) are chosen to have random direction and normal distribution with zero mean and standard deviation given by the average velocity calculated from the previous step.

Finally, the number of clusters is defined as the number of maxima in the potential energy distribution, averaged over all five repetitions. Points with a positive potential energy are classified as outliers. The rest is split into clusters according to the particles’ potential energy level.

2.2. Data

A series of test datasets with numbers of clusters and cluster members, as well as shape and dimensionality were generated. Here, three toy problems are presented: set 1 consisting of a ‘horseshoe’ and two spheroids overlaid by random noise in 3-dimensional space (461, 36, 79 and 57 points, respectively, Figure 1(a)); set 2 consisting of two dense regions positioned around two skew lines overlaid by random noise in 3-dimensional space (200, 320 and 100 points, respectively; Figure 1(b)); and set 3 consisting of two dense regions positioned around two skew lines overlaid by random noise in 100-dimensional space (70, 90 and 35 points, no graphic depiction possible).

![Figure 1.](image)

In addition, data from 43 MS patients (17 male and 26 female subjects with a mean age of 41±11 years), scanned on a 3T TRIO Siemens scanner, were analyzed. The expanded disability status scale (EDSS) score [4] of the patients at the time of the examination ranged between 1 and 6.5. Quantitative \( T_1^* \), \( T_2^* \), and absolute water content maps were constructed for each subject as described in [6]. The individual maps were spatially normalized to the MNI ICBM452 brain template, so that same anatomical structures appear at the same coordinates in the maps. From the reconstructed maps, a 494-dimensional feature vectors was calculated for each subject. The features include the average and standard deviation of all quantitative MR parameters for the whole brain as well as for each individual slice. Furthermore, the parameter histograms were included as additional features in order to implicitly include higher moments of the respective distributions.

3. Results

Applying the Nuclear Potential Clustering on the test datasets leads to a correct clustering of 96% of the points of set 1, 99% of the points of set 2 and 98% of the points of set 3. Figure 2 demonstrates the final data distribution for the ‘horseshoe’ test set. As it can be seen in Figure 2(a), nearby nucleons fuse to almost point-like nuclei. Figure 2(b) depicts the corresponding final energy distribution. The different horizontal lines represent the different clusters, as all nucleons in a point-like spherical nucleus have, to a first approximation, the same potential energy. This distribution is used to define the number of groups formed and to assign the points to the appropriate clusters. The corresponding result is shown in Figure 2(c). Similar figures for the second and the third test data sets as well as videos demonstrating their spatial (for the 3D data) and energy evolution can be found at:

http://www.rheinahrckampus.de/Nuclear-Potential-Clustering.5142.0.html
Applying the algorithm on the MS data resulted in a splitting of all points into two main categories. The first group consists of 27 subjects with an average EDSS of 0.67±1.58 and an average age of 37±10 years. The second group consists of 16 points classified as noise. It contains primarily older patients with a mean age of 46±11 years and a higher EDSS score of 2.50±2.29.

4. Discussion

Nuclear Potential Clustering showed a good performance on all artificial data sets examined. Taking a closer look at our results showed that the falsely classified data are mostly particles on the rim of a cluster where the nucleon density decreases, or noise that is accidentally positioned nearby or inside of a denser group. It is yet important to note that the algorithm is sensitive to variations of R. Omitting the modulation of the short-range strong force leads to building of sub-formations depending on the intra-cluster density variation. However this could even be advantageous, if one is interested in exploring the inner structure of a larger group. Our goal here was to detect major clusters, which was successfully achieved by applying the parameters described in section 2.1.

Analyzing the MS data showed that patients with low grade disease (mean EDSS=0.67) form a homogeneous group, meaning that their image derived features are comparable. On the other hand, patients with a higher EDSS grade (mean EDSS=2.7) are identified as noise. This appears reasonable as higher disease grades are expected to be much more heterogeneous. Indeed our data reveal that this group does not show a consistent pattern, at least with respect to features derived from quantitative $T_1^*$, $T_2^*$, and absolute water content maps.

In summary, we have developed a new method for grouping high dimensional data based on a simplified model of nuclear fusion. The new approach allows for the clustering of non-spherical, correlated datasets, which are typically observed in MR derived image features. It therefore presents an ideal tool for the automated analysis of medical imaging data, not only in the MRI domain.

References

[1] Duda RO, Hart PE, Stork DG 2001, *Pattern Classification* (New York: Wiley-Interscience)
[2] Neema M, Stankiewicz J, Arora A, Guss ZD, Bakshi R. MRI in MultipleSclerosis: What'sInside the Toolbox?, Neurotherapeutics 2007;4: 602-617.
[3] Bakshi R, Thompson AJ, Rocca MA, Pelletier D, Doussset V, Barkhof F, Inglese M, Guttmann CRG, Horsfield MA, Filippi M. MRI in multiple sclerosis: current status and future prospects. Lancet Neurol 2008; 7:615-625.
[4] Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 1983; 33 (11): 1444–52.
[5] Krane KS 1988, *Introductory Nuclear Physics* (New York: Wiley & Sons)
[6] Neeb H Ermer V, Stoecker T, Zilles K, Shah NJ. Rapid high-resolution mapping of cerebral water content with full brain coverage. Neuroimage 2008; 42(3): 1094-1109.
[7] Woods RD, Saxon DS. Diffuse surface optical model for nucleon-nuclei scattering. Phys. Rev. 1954; 95: 577–578