Nature Inspired Learning Models

Dymitr Ruta* and Bogdan Gabrys†

ABSTRACT
Intelligent learning mechanisms found in natural world remain are still in their learning performance and efficiency of dealing with uncertain information coming in a variety of forms, yet remain under continuous challenge from human driven artificial intelligence methods. This work intends to demonstrate how the phenomena observed in physical world can be directly used to guide artificial learning models. An inspiration for the new learning methods has been found in the mechanics of physical fields found in both micro and macro scale. Exploiting the analogies between data and particles subjected to gravity, electrostatic and gas particle fields, new algorithms have been developed and applied to classification and clustering while the properties of the field further reused in regression and visualisation of classification and classifier fusion. The paper covers extensive pictorial examples and visual interpretations of the presented techniques along with some testing over the well-known real and artificial datasets, compared when possible to the traditional methods.

KEYWORDS
Machine learning, classification, classifier fusion, clustering, regression, visualisation, gravitational field, electrostatic field, Lennard-Jones potential

1 Physics of information

It is well known that information has clear ties with physical world. Every single item from the physical world holds enormous amounts of information, we humans possibly barely know of. It is even believed that the reality of material world arises at the very bottom from elementary yes-no questions capturing single bits of information [1]. The comprehensive success of computing seems to support this concept of *it from bit*. Despite the natural familiarity with information, there is still no uniform definition covering all aspects of information. So far the most advanced theory of information assumes information entropy as a probabilistic measure of information content [2]. Inspired by the analogy to the thermodynamic entropy it turned out that the entropy of information effectively measures the uncertainty arising from the ambiguity of competitive choices. Vagueness perceived as inability of making sharp distinctions in the world is another type of uncertainty, which due to crisp perception of evidence, probabilistic models cannot capture [2]. The measure of fuzziness is just an example of this new dimension of uncertainty arising from the fuzzy set theory [2]. Further investigations conducted within mathematical theory of evidence revealed even more new dimensions of uncertainty [2]. Analogy between information uncertainty and physical energy seems to be more than tempting. Similarly to energy appearing in a variety of forms, there are many different types of uncertainty. Like for the energy measured in Jules, all uncertainty measures yield values in the same units of logical bits. Moreover, energy viewed in general as a capacity for doing work corresponds to uncertainty appearing as a capacity for obtaining information.

Pushing this analogy a step further one can expect similar relationship between data and matter. Like for the matter being a stable representation of energy, the data points are the true unstructured embodiments of information. Such inspirations are not unique in research. Already mentioned Shannon entropy representing probabilistic interpretation of data randomness is an example of a direct counterpart to the thermodynamic entropy describing the degree of order among physical gas particles. Ongoing advances in quantum information theory show an excellent example where the information bits converge with the elementary matter units [3]. The mathematical concept of a field, so commonly observed in nature, has hardly been exploited in the pattern recognition domain. In [4] Hochreiter and Mozer use electric field metaphor to Independent Component Analysis (ICA) problem where joint and factorial density estimates are treated as a distribution of positive and negative charges. In [5] Principe et al introduces the concept of information potentials and forces employing unconventional definition of mutual information based on Renyi's entropy. Torkkola used further these concepts for linear [6] and non-linear [7] transformations of the data maximising their mutual information.

*Intelligence Systems Research Centre, British Telecom Group, Research & Venturing, Orion Building 1st floor, pp12, Adastral Park, Martlesham Heath, Ipswich IP5 3RE, UK, dymitr.ruta@bt.com

†Computational Intelligence Research Group, Bournemouth University, School of Design, Engineering & Computing, Poole House, Talbot Campus, Fern Barrow Poole BH12 5BB, United Kingdom, bgabrys@bournemouth.ac.uk

This work intends to demonstrate how the phenomena observed in physical world can be directly used to guide artificial learning process. In our approach inspiration for the new learning method has been found in the mechanics of the potential fields found in both micro and macro scale. Across the artificial learning domain new field based algorithms have been developed and applied to classification and clustering while the properties of the field further reused in regression and visualisation of classification and classifier fusion. In all these methods the data points were considered as particles that carry elementary units of charge generating a central field that acts upon other samples in the input space.

Guided by the mechanics of the gravitational field we devised a new clustering method in which data points embodied by mass particles became the sources of the attracting field naturally grouping the data up to the ultimate collapse into a single cluster. Further refinement of such clustering method has been found following gas particle dynamics defined by the Lennard-Jones potential. Here the field, like for gas particles, has dual nature of attracting and repelling depending on the distance between sources, and for the clustering purposes it has been made attracting on short distances and repelling on longer distances to encouraging better cluster separation.

Gravity field was also the starting point for the supervised learning or classification. Unlike for clustering, the labelled training patterns were static, kept affixed in the input space while the testing samples were let free in the input space to ultimately meet one of the training pattern and share its label. In order to exploit the label information coming as extra information on top of spatial data distribution, a refinement was proposed that directly models the electrostatic field, yet allowing as many different types of source charges as many different classes are in the dataset. As a result of this refinement a repelling force was devised as a field action between differently labelled data which again encouraged better class separation and smoother class boundaries.

The field concept has been used also in other learning aspects. In the regression domain, the field was used in a sense of building an interpolated regression surface out of the training samples. In the visualisation theme, a field approach was used to show discriminant functions of various classifiers and advanced to the level which allows to fully visualise not only classifiers but also classifier fusion methods.

The paper covers extensive pictorial examples and visual interpretations of the presented techniques. The practical applicability of some model is examined and tested on the well-known real and artificial datasets and compared when possible to the existing techniques.

The remainder of the paper is organized as follows. Section 2 explains gravity and electrostatic field inspired classification models. The next section provides details of the clustering algorithms build upon principles in gravity field and particle dynamics in noble gases. Section 4 explains how field can be reused in the regression based on data interpolation and the following section presents interesting classification visualisation methodology again with the aid of underlying field concept. In Section some results from experiments carried out on real dataset are presented followed with the concluding remarks briefly presented in the closing section.

2 Classification field models

The concept of a field in classification is not new and in fact is related to the kernel methods [8]. The rationale behind using the field concept is to ensure that every data sample is actively contributing in the formation of final classification decision. All the data are considered to be a charged particles each being the source of a central field affecting other samples. All the characteristics of such field are the results of the definition of potential and can be absolutely arbitrarily chosen depending on various priorities. For classification purposes the idea is to assign the class label to a previously unseen sample based on the class spatial topology learnt from the training data. This goal is achievable within the data field framework if we assume testing samples to be mobile and forced by the field to move towards affixed training data to share their label. The overall field measured in a particular point of the input space is a result of a superposition of the local fields coming from all the sources. Thus the positions of the training data uniquely determine the field in the whole input space and by that determine the trajectories of the testing data during classification process. If the field is designed in such a way that all trajectories possible end up in one of the sources then the whole feature space can be partitioned into regions representing distinct classes. The boundaries between these regions form the ultimate class decision boundaries, which completes the classifier design process.

2.1 Attracting gravity field model

Inspired by the field properties of the physical world one can consider each data point as a source of a certain field affecting the other data in the input space. In general the choice of a field definition is virtually unrestricted. However, for the classification purposes considered in this paper, we use a central field with a negative potential increasing with the distance from a source. An example of such a field is the omnipresent gravity field. Given the training data acting as field sources, every point of the input space can be uniquely described by the field properties measured as a superposition of the influences from all field sources. In this paper we consider a static

field in a sense that the field sources are fixed to their initial positions in the input space. All dynamic aspects imposed by the field are ignored in this work. Given a training set of n data points: $X = (\mathbf{x_1}, .., \mathbf{x_n})$ let let each sample be the source of a field defined by a potential $U_j = -cs_i f(\overrightarrow{r_{ij}})$ where c represents the field constant, s_i stands for the source charge of $\mathbf{x_i}$, and $f(\overrightarrow{r_{ij}})$ is a certain non-negative function decreasing with an increasing distance $|\overrightarrow{r_{ij}}| = |\mathbf{r_{ij}}|$ from the source $\mathbf{x_i}$ to the point $\mathbf{y_j}$ in the input space. In the gravitational field we simply have $f(\mathbf{r_{ij}}) = 1/|\mathbf{r_{ij}}|$. Overall the potential U_j and field interaction energy E_j in a certain point $\mathbf{y_j}$ of the input space is a superposition of the potentials coming from all the sources:

$$U_j = -c\sum_{i=1}^{n} \frac{s_i}{|\mathbf{r_{ij}}|} \qquad E_j = -cs_j \sum_{i=1}^{n} \frac{s_i}{|\mathbf{r_{ij}}|} \tag{1}$$

We can simplify model further by assuming that all data points are equally important and have the same charge equal to the unit: $s_i = 1$ thus eliminating it from the equations 1. Another crucial field property is its intensity E_j, which is simply a gradient of the potential and its solution leads to the following:

$$\overrightarrow{E_j} = \mathbf{E_j} = -\overrightarrow{\nabla} U_j = -\left(\frac{\partial U_j}{\partial y_{j1}}, ..., \frac{\partial U_j}{\partial y_{jm}}\right) = -c\left(\sum_{i=1}^{n} \frac{y_{j1} - x_{i1}}{|\mathbf{r_{ij}}|^3}, .., \sum_{i=1}^{n} \frac{y_{j1} - x_{i1}}{|\mathbf{r_{ij}}|^3}\right) = -c\sum_{i=1}^{n} \frac{\mathbf{y_j} - \mathbf{x_i}}{|\mathbf{r_{ij}}|^3} \tag{2}$$

A field vector shows the direction and the magnitude of the maximum decrease in field potential. By further analogy to gravitational field, the charged data point is affected by the field in the form of force attempting to move the sample towards lowest energy levels. As the charge has been assumed uniformly of unit value and excluded from the equations the force vector becomes identical to field intensity: $\overrightarrow{F_j} = \mathbf{F_j} = s_j \mathbf{E_j} = \mathbf{E_j}$.

The concept of the field forces will be directly exploited for the classification process. The field constant c does not affect the directions of forces but only decides about their magnitudes, hence without any loss of generality we can assume its value as unit and that way free the field equations from any parameters, apart from the definition of the distance itself.

The labelled training data uniquely determine the field and all its properties in the whole input space, yet there is no training process per-say other than just acknowledging the training data as field sources. All the calculations required to classify new data are carried out online during the classification process. Computationally the critical process is the calculation of the distances from the examined points and all the sources. Using matrix formulation of the problem and the appropriate mathematical software, this task can be obtained rapidly even for thousands of training sources. Denoting by $Y^{[N\times m]}$ a matrix of N m-dimensional data points to be classified and by $X^{[n\times m]}$ the matrix of n training data, the task is to obtain the matrix D of all the distances between the examined data and the training data. Introducing "$\circ$" as element-wise matrix multiplication and $\mathbf{1}^{[n\times m]}$ as an n by m matrix with all unit elements, the distance matrix can be calculated instantly by:

$$D = Y \circ Y \bullet \mathbf{1}^{[1\times n]} - 2 \bullet Y \bullet X^T + \mathbf{1}^{[N\times 1]} \bullet X \circ X \tag{3}$$

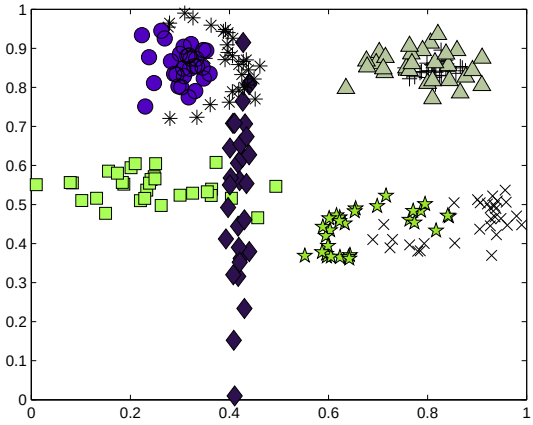
Given the distance matrix the classification process is very straightforward. The data points to be classified are simply placed in the input space and let to slide down the potential wells to meet one of the field source and share its label. Ignoring the dynamics of sliding data i.e. removing the kinetic energy they gain during descending towards field sources classification can be organised into a step-wise process at which testing data is moved by $\triangle Y^{[N\times m]}$ again efficiently calculated using matrix formulation by:

$$\triangle Y = \frac{d \bullet \mathbf{F}}{(F \circ F) \bullet \mathbf{1}^{[m\times 1]} \bullet \mathbf{1}^{[1\times m]}} \tag{4}$$

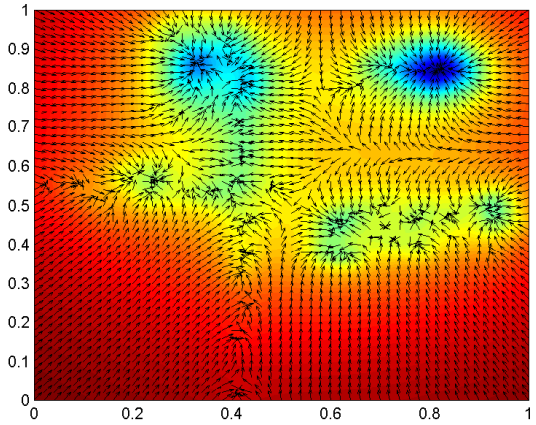
where d is an arbitrarily small step. Negative potential definition (1) ensures that testing data will always meet one of the field source. To avoid numerical problems the data should be normalised within the same limits and distances limited from the bottom by a small interception threshold comparable to d which prevents division by zero and "overshooting" the field source. We call such model the gravity field classifer (GFC). Figure 1 demonstrates the classification process with the GFC classifier applied to the artificial dataset of 240 2-D samples with 8 classes.

2.2 Repelling electrostatic field model

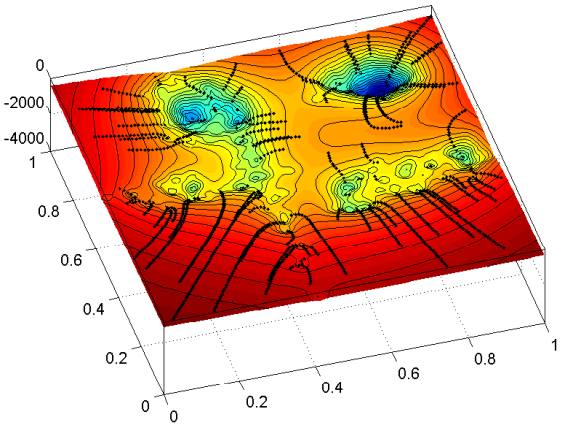
So far data have been only attracted to each other which was a consequence of the negative potential definition resulting in the energy wells intercepting samples found in the neighbourhood. Such field does not use the information about the class labels as all the samples are considered to hold the same charge. Ideally, the attracting force should be acting only upon the data from the same class. At the same time the samples from different classes should be repelled from each other to stimulate increased separability between classes. Again



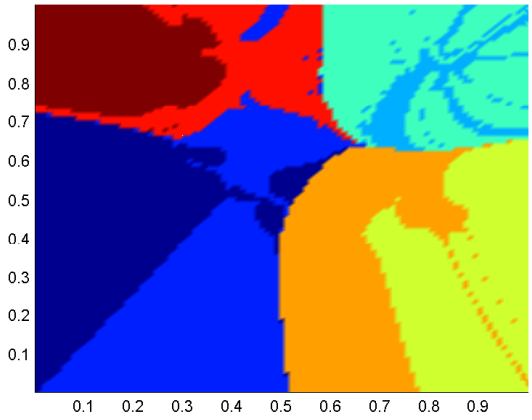
(a) Dataset: 240 2-D 8-class



(b) Vector field plot



(c) Classification process



(d) Resulting class boundaries

Figure 1: Visualisation of the static gravity field based classification process performed on the 8-classes 2-D artificial data of 240 samples. 1(a) Dataset plot. 1(b) Vector plot of resulting field intensities. 1(c) 3D visualisation of the classification process using GFC. 1(d) Class boundaries generated by the GFC classifier

nature offers a perfect guide in a form of electrostatic field, where opposite charges attract each other and charges of the same sign repel from each other. To adopt this rule to the labelled data, the samples from the same class should interact with negative potential as in previous case, whereas samples from different classes should generate the positive potential of the same absolute value triggering repelling force.

The major problem with electrostatic data field is that testing samples do not have labels and cannot straightforwardly interact with labelled training samples. Estimating the label of the testing sample means that classification is accomplished. To avoid this trivial situation we assume that each testing sample is decomposed into fractional subsamples labelled by all classes present in the field. Charges of such subsamples are proportional to the class potentials and normalised to sum up to a unit.

Given the training set $X = (\mathbf{x_1}, .., \mathbf{x_n})$ with labels $L_S = (l_1, .., l_n)$ where $l_i \in (1, .., C)$, labels partition matrix $P^{N \times C}$ for the testing set $Y = (\mathbf{y_1}, .., \mathbf{y_N})$ can be simply obtained by $p_{jk} = |U_j^k| / \sum_{i=1}^{C} |U_j^i|$ where U_a^b stands for potential generated by samples $\mathbf{x_i}$ coming from class indexed by b measured in point $\mathbf{y_a}$. Given this label partition matrix the new definition of potential and field vector take the following form:

$$U_j = \sum_{i=1}^{n} \left(\underbrace{\frac{\sum_{k \neq l_i} p_{jk}}{|\mathbf{r_{ij}}|}}_{repelling} - \underbrace{\frac{p_{jl_i}}{|\mathbf{r_{ij}}|}}_{attraction} \right) = \sum_{i=1}^{n} \frac{1 - 2p_{jl_i}}{|\mathbf{r_{ij}}|} \qquad \mathbf{E_j} = \sum_{i=1} n \left[(1 - 2p_{jl_i}) \frac{\mathbf{y_j} - \mathbf{x_i}}{|\mathbf{r_{ij}}|^3} \right] \tag{5}$$

The numerator of the potential definition (5) can be both positive and negative depending on the class partial memberships. In the presence of many classes, regardless of their topology, the absolute values of the partition matrix P will naturally decrease to share the evidence among many classes. Effectively the potential would grow positive with the repelling force dominating the field landscape.

In our model the data still have to slide down the potential towards the source samples. To satisfy this it is sufficient to normalise the field such that the overall potential of the whole field should not be larger than zero. Taking into account the fact that the field is substantially negative in the close neighbourhoods around the training samples, it is sufficient to satisfy the condition of $\sum_{j=0}^{N} U_j$. To achieve this goal potential definition has to be parameterised and solved with respect to the regularisation coefficient q as in the following:

$$\sum_{j=1}^{N} U_j = \sum_{j=1}^{N} \sum_{i-1}^{n} \frac{1 - qp_{jl_i}}{|\mathbf{r_{ij}}|} \tag{6}$$

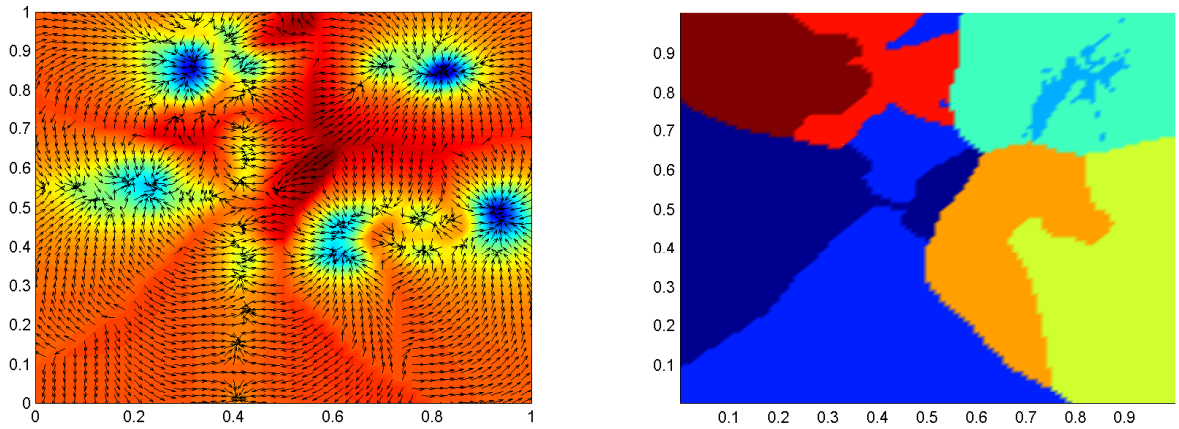
In the model we use bisection method to find numerical estimation of the parameter q. Note that parameter q has a meaningful interpretation as the value $1 - q$ says in general how many times should the attractive interaction be stronger to compensate the excess of the repelling interaction coming from the multitude of different classes. Having met all conditions discussed above the classification process follows the same routine as in the gravity model, and this time due to direct inspiration from physical electrostatic field we will refer to the presented method as electrostatic field classifier or shortly EFC. Example of such field is presented in Figure 2 again showing the whole classification process applied to the same dataset as in Figure 1.

3 Clustering field models

In clustering the objective is to group the data objects into a set of disjoint classes, called clusters, such that the objects within a class have high similarity to each other, while objects in separate classes are more dissimilar. Clustering is an example of unsupervised learning in which all the data objects are automatically assigned to a set of distinct classes, without using any predefined information about the classes or class structure. One of the greatest problems in clustering is inability to formally justify the best number of clusters [9]. As a result hierarchical clustering methods with a support of the dendrogram remain the most common as they have a very flexible mechanism of delivering the clustering at any granularity level starting from n singleton data clusters up to the single cluster. In the continuous numerical space considered in this work, the most convenient similarity measure is simply an Euclidean distance and again attracting properties of the physical fields can be reused in the attempt to construct a field based hierarchical clustering algorithms.

3.1 Hierarchical gravity based clustering model

In the simplest form we reuse again gravity field model to simulate self collapse of the data-particles as in the case of attracting gravity field classifier. The only difference in the case of clustering is that the field is not static and the field sources move along the field forces. Whenever two or more sources approach each other at the arbitrarily small distance d they are considered a single yet *heavier* clusters which continuous to move according to gravity mass dynamics equations. As before the kinetic energy that moving samples gain after each



(a) Vector field plot

(b) Resulting class boundaries

Figure 2: Visualisation of the electrostatic field based classification process performed on the 8-classes 2-D artificial data of 240 samples. 2(a) Vector plot of resulting field intensities. 2(b) Class boundaries generated by the EFC classifier

step is cancelled such that the simulation has a character of converged all-to-one collapse rather than energy preserving elliptical cycles. During the collapse the place of cluster merging is irrelevant. The only thing that is important is the timing of mergers and what samples took part in the mergers. The stepwise sequence of such mergers uniquely determines the clustering dendrogram and thereby completes the hierarchical clustering in a single cluster around the data mean.

The process of field and forces calculation is here identical to the GFC classifier shown in (1)-(4) with the only difference that the field charge multiplier s_i equal to the number of original data points in the i^{th} cluster can not be made redundant as it varies and needs recalculation at each step. Figure 3 shows how the 8-classes 2-D dataset from Figure 1(a) as well as well known Iris dataset groups here into converging clusters at different point of time, which could correspond to the dynamic equivalent of dendrogram.

3.2 Lennard-Jones potential model

Due to the centrical nature of the gravity field clusters constructed as a result of its action tend to be elongated radially from the data mean point. To introduce an element of resistance to this behaviour, attracting-only gravity field can be replaced by the field that attracts locally and repels distant samples from each other similarly to the electrostatic field classifier discussed in Section 2.2. Here again the inspirations have been found in nature. Namely, copying gas molecules dynamics ruled by the Lennard-Jones potential [10]. In noble gasses the interaction potential between a pair of molecules is given by:

$$U_{ij} = 4\epsilon \left[(\sigma/\mathbf{r_{ij}})^{12} - (\sigma/\mathbf{r_{ij}})^{6} \right] \tag{7}$$

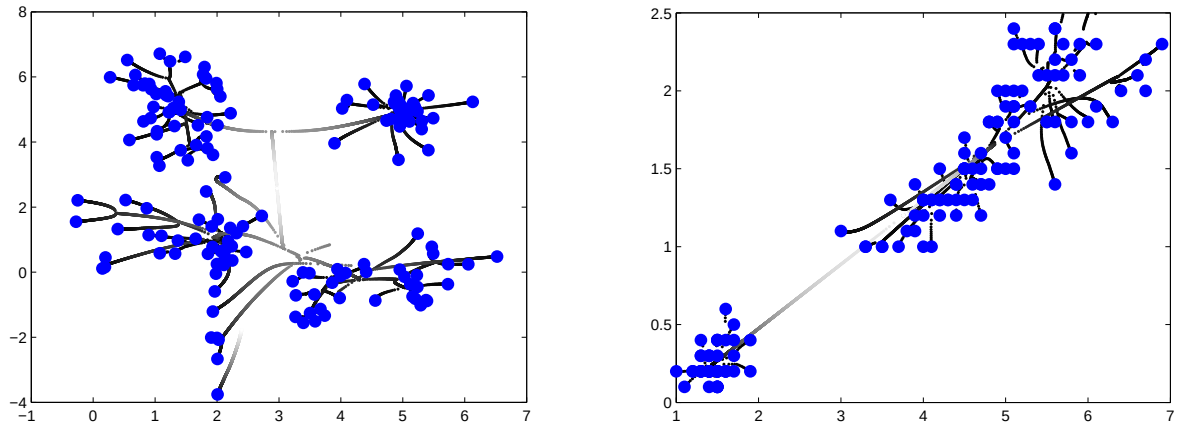
If all n data points are considered to be such interacting field sources then each of them will be subjected to the field intensity vector:

$$\overrightarrow{E_j} = \mathbf{E_j} = -\overrightarrow{\nabla} U_j = -\left(\frac{\partial U_j}{\partial y_{j1}}, ..., \frac{\partial U_j}{\partial y_{jm}} \right) = \frac{24\epsilon \sum_{i=1}^{n} s_i s_j (\mathbf{y_j} - \mathbf{x_i})}{|\mathbf{r_{ij}}^2|} \left[\left(\frac{\sigma}{|\mathbf{r_{ij}}|} \right)^6 - 2 \left(\frac{\sigma}{|\mathbf{r_{ij}}|} \right)^{12} \right] \tag{8}$$

The force being identical to the field intensity vector has been simplified to the following form:

$$\mathbf{F_j} = c \sum_{i=1}^{n} s_i s_j (\mathbf{y_j} - \mathbf{x_i}) \left[\left(\frac{\sigma}{|\mathbf{r_{ij}}|} \right)^a - \left(\frac{\sigma}{|\mathbf{r_{ij}}|} \right)^b \right] \tag{9}$$

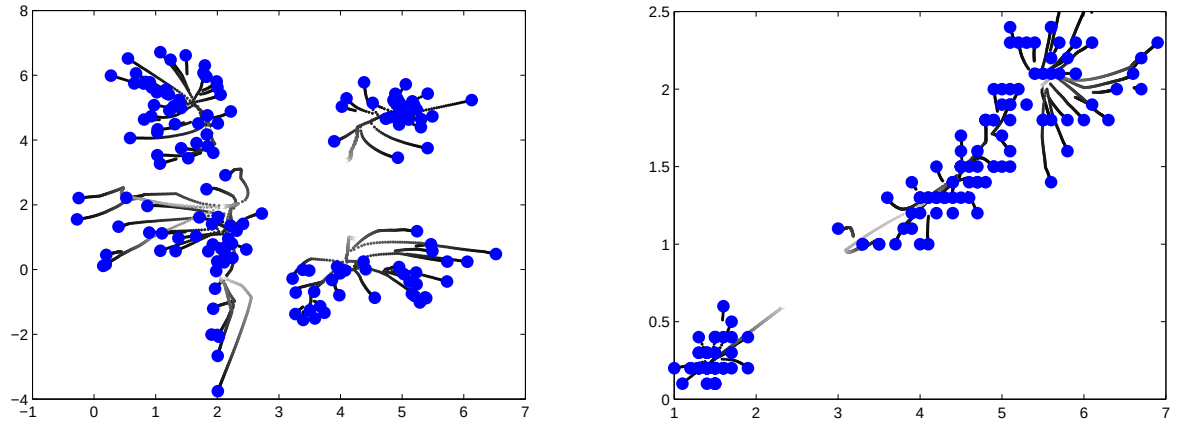
which retains its character yet is simpler to interpret and handle computationally for lower powers of a and b assuming $a < b$. From (9) it is clear that σ is the distance at which interaction potential disappears and there is no force acting upon the data. For distances greater than σ the field exerts attracting force while for distances smaller than σ the force changes the sign to positive and becomes repelling. Greater flexibility in designing the



(a) 8 class 2D dataset - dynamic dendrogram

(b) Iris dataset - dynamic dendrogram

Figure 3: Visualisation of the hierarchical clustering using gravity field. Lightening trajectories correspond to the passing time



(a) 8 class 2D dataset - dynamic dendrogram

(b) Iris dataset - dynamic dendrogram

Figure 4: Visualisation of the hierarchical clustering using gas model based on Lennard-Jones potential with the repelling barrier in the rage of 4(a) 3-6 and 4(b) 1-3

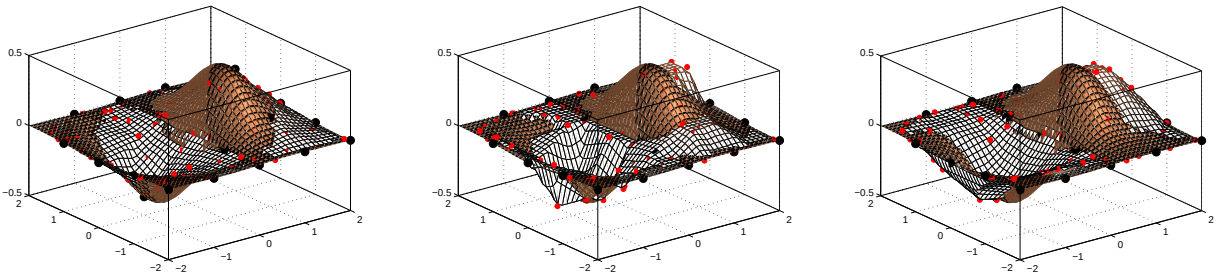
balance between repelling and attracting field can be obtained by enforcing exclusiv potential definitions on different distance ranges. In our final gas model we define the following field:

$$\mathbf{F_j} = c\sum_{i=1}^{n} s_i s_j(\mathbf{y_j} - \mathbf{x_i})\left[\left(\frac{(r > \sigma_1}{|\mathbf{r_{ij}}|}\right)^a - \left(\frac{\sigma_1 \leq r \leq \sigma_2}{|\mathbf{r_{ij}}|}\right)^b + \left(\frac{r < \sigma_2}{|\mathbf{r_{ij}}|}\right)^a\right] \tag{10}$$

where field is attracting in whole distance range excluding the range $(\sigma_1 : \sigma_2)$ where the field is repelling. The logical operators appearing in the numerators in 10 are assumed to return 1 if are true and 0 otherwise. Figure 4 shows an example of such model with the parameters $a = 1$ and $b = 2$. Note that for Iris dataset the presented clustering model correctly grouped dataset into 3 classes, despite the fact that the two classes on the right are quite attached to each other. As a rule of the thumb we assumed the width of the potential barrier to be a third of the average data spread.

4 Interpolation based regression model

In the regression domain the task is simply to find the mapping between continuous input and output given some observations of $\mathbf{x} \rightarrow y$ used to train the regression model. There is many ways these task can be accomplished



(a) Linear interpolation (b) Nearest neighbour interpolation (c) Spline interpolation

Figure 5: Example of various interpolation surfaces generated on a set of 100 random points (smaller red dots), based on grid training set (large black dots) and overlayed on the true regression surface.

from simple linear regression up to the complex non-linear methods like neural networks [9]. Staying within the field concept we consider the outputs of the training data as samples of the underlying field which is generated by the input data. The assumption is that the field spreads along the whole space and the more input-output pairs is available the more of the output field is revealed. From this perspective regression can be represented by a simple interpolation that tries to find the location on the output field based on the neighbouring training samples. There is a reach choice of interpolation methods. For a simple demonstration let us consider a simple 2-input to 1-output function $z(x, y) = x(-x^2 - y^2)$. Let the training set be a regular unit grid in the range of $(-2 : 2)$ for both x and y and the output values are calculated according to the introduced function $z(x, y)$. A testing set of 100 input data points has been randomly sampled from the considered range and its output obtained from interpolation based on the given grid of training data. Figure 5 shows the plots of function $z(x, y)$ and the interpolation surface aroused on the testing points.

5 Field based classification visualisation

In business applications the most valuable classifiers are those that work transparently such that the user can control and understand the decisional process at all stages. Not surprisingly, a simple decision tree is by far the most popular classifier used in commercial applications as it allows for a step-by-step monitoring and understanding of the classification decision. It is psychologically proven that visuals transmit much more of conscience packed information than text and are also much easier to be remembered. Surprisingly there is not much evidence of this advantage being taken in commercial applications of pattern recognition. Many visualisation techniques have been proposed in the sister domain of knowledge discovery [11], yet even huge analytical packages like Mirage [12] miss out a comprehensive model snapshot that would visually explain how classification models are generated, such that the user would understand the link between the data, the model is built on, and the model itself. The field concept has a lot to offer in terms of model visualisation capability. If we assume that instead of charge, as before, the data points are the sources of probabilistic evidence, than the fields aroused on the labelled datasets become posterior class probability maps, usually called discriminant functions and constitute the whole internal description of the classification model. Formally, given the feature vectors $\mathbf{x} \in X$, the objective of a classifier, is to assign the new pattern $\mathbf{x}$ to a relevant class $\omega_j \subset \Omega$, where $\Omega = \{\omega_1, ..., \omega_C\}$, based on previous observations of labelled patterns: $X_T = \{\mathbf{x}, \omega\}$. The classification model takes the form of a set of discriminant functions $g_j(\mathbf{x}) = P(\omega_j \mid \mathbf{x})$ calculated separately for all classes. Examples of such discriminant function plots are shown for decision tree and quadratic classifier in Figure 6. The final classification decision is formulated as:

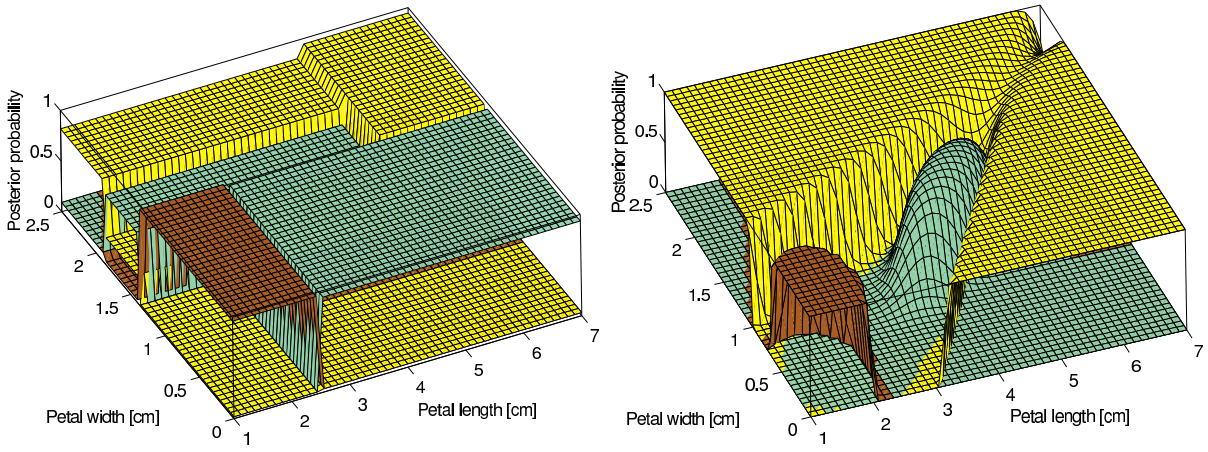
$$\omega_d = \arg \max_{j=1}^{C} P(\omega_j \mid \mathbf{x}) \tag{11}$$

which visually means that the class corresponding to the top most discriminant surface becomes the classifier decision for particular data point. Let the decision surface be defined by:

$$D = \max_{j=1}^{C} P(\omega_j \mid \mathbf{x}) \tag{12}$$

Decision surface can be easily picked up from Figure 6, where it becomes simply the superposition of the top most patches of individual class discriminant surfaces.

Using discriminant functions as internal model description, the fusion of classifiers becomes a straightforward process. It involves fusing individual discriminant functions into a combined discriminant function for each of



(a) Decision tree (b) Quadratic classifier

Figure 6: Visualisation of the discriminant functions applied to the 3-class Iris dataset. 6(a) Discriminant functions generated by the decision tree classifier. 6(b) Quadratic classifier discriminant functions

the classes and then selecting the class with the maximum combined posterior probability. There are many fusion operators by means of which classifiers can be combined and all of them can be easily visualised by means of plots of discriminant functions and ultimately combined decision surfaces.

Given a pool of N classifiers let $P_{ij} = P_i(\omega_j \mid \mathbf{x})$ stand for the discriminant function of the i^{th} $(i = 1, .., N)$ classifier applied to the class ω_j $(j = 1, .., C)$. The major classifier fusion operators [13] can easily be expressed by simple aggregation operations carried out on the discriminant functions P_{ij}. Note that all combiners first combine the discriminant functions for all classes and by doing so define a new composite classifier which has its own set of combined discriminant functions. That means that one can similarly visualise the classifier fusion and likewise define its decision surface by:

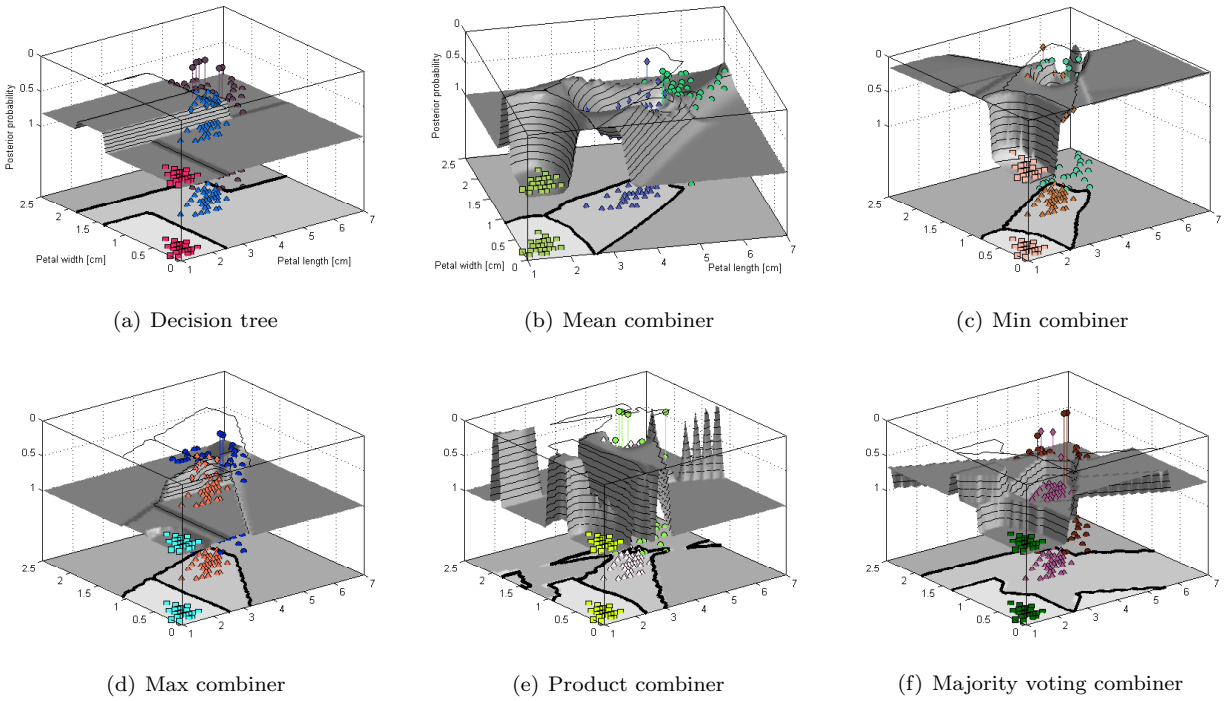
$$D_F = \max_{j=1}^{C} F_{i=1}^{N} P_{ij} \tag{13}$$

where F denotes the specific fusion operator.

Although fusion of discriminant functions is scalable to many dimensions, the visualisation of it is meaningful only in 2 or 3 dimensions which can be observed by the user. The most informative visualisation model would capture not only the shape of the decision surface and the corresponding decision boundaries but also would illustrate the data samples and their link to the process of the creation of discriminant functions. To achieve this goal the decision surface is plotted upside down such that the observer looking from above can penetrate the regions of lower posterior probability and hence observe misclassified samples. Figure 7(a) shows an example of the complete visualisation of the decision tree classifier, while Figures 7(b)-7(f) illustrate the fusion of the decision tree, quadratic and neural network classifiers by means of selected fusion operators. Decision surface is plotted in the top half of the presentation cube such that one can clearly see the decision boundaries plot in the bottom plane and understand where the boundaries are coming from. Note that the data points which do not lie on the decision surface represent samples which are misclassified by the classifier or combiner. Moreover their distance from the decision surface along vertical axis indicates how much more of posterior probability was allocated by the model to the incorrect class compared to the true class, probability of which can be read from the height of the data points.

6 Experiments

Due to space restrictions experimentation has been narrowed down to the performance tests of presented classification models. Both GFC and EFC classifier have been tested over 6 dataset along with 15 other classifiers for comparison. Figure 8 shows the details of datasets and classifiers used in this experiment. For the first 4 datasets we applied random splitting into two equal parts used for training and testing respectively. For the Segment and Shuttle dataset, due to their large sizes we used 10-fold cross-validation for performance estimation. Table 1 shows averaged individual performances of classifiers from this experiment. Although the performances of GFC and EFC classifiers is never the best they consistently remain among the best classifiers. This attribute makes it an interesting alternative to other classifiers that could be particularly useful in combining classifiers.



(a) Decision tree (b) Mean combiner (c) Min combiner

(d) Max combiner (e) Product combiner (f) Majority voting combiner

Figure 7: Complete visualisation of the decision tree classifier and mean combiner built on decision tree, quadratic and neural net classifiers applied to iris dataset.

Dataset	samples	features	classes
Iris	150	4	3
Conetorus	400	2	3
Gaussians	250	2	2
Azizah	291	8	20
Segment	2310	19	7
Shuttle	58000	9	7

(a) Datasets

Classifier	Description
klclc	PCA based linear classifier
loglc	logistic regression based classifier
fisherc	Minimum least squares classifier
ldc	Linear normal density based classifier
nmc	Nearest mean linear classifier
qdc	Quadratic normal density based classifier
qua	Quadratic classifier
svc	Support vector machine clasifier
knn	k nearest neighbour classifier
parzenc	Parzen density estimation classifier
subsc	Random subspace classifier
treec	Decision tree classifier
lmnc	Feed-forward neural net clasdifier
rbnc	Radial basis neural net classifier
bpxnc	bpxnc - Back-propagation neural net classifier

(b) Classifiers

Figure 8: A list of 6 datasets (8(a)) and 15 classifiers (8(b)) used in experiments for performance comparison with GFC and EFC classifiers.

Table 1: 10-fold cross-validation error rates [%] obtained for 17 classifiers and 6 datasets described in Table iments.

	klc	log	svc	ldc	nmc	qdc	qua	svc	knn	par	sub	tre	lmn	rbn	bpn	gfc	efc
Iri	2.47	4.72	3.39	2.43	7.97	2.53	3.23	4.16	4.49	4.25	2.85	5.85	4.59	11.28	4.28	4.56	4.33
Cnt	27.35	25.58	26.54	27.23	29.44	20.32	19.86	17.05	16.25	60.02	18.78	18.60	31.83	18.98	18.98	18.89	16.69
Gau	15.17	14.44	15.17	14.71	28.71	15.11	15.11	14.74	13.26	13.17	23.55	18.64	14.09	17.71	13.91	15.97	14.25
Azi	41.62	50.65	45.39	32.66	45.58	89.94	98.70	28.70	31.17	35.78	49.74	80.39	46.49	53.90	37.73	58.44	32.34
Seg	18.37	19.21	15.59	17.82	15.79	14.85	17.28	6.34	9.60	10.00	18.27	36.44	16.83	16.98	10.10	13.37	10.54
Shu	16.95	12.95	15.45	10.35	34.50	10.45	9.74	3.55	4.55	9.80	35.75	11.85	13.20	40.25	11.20	4.60	6.50

7 Conclusions

In this paper, we demonstrated that a number of natural physical phenomena can be directly used to devise an artificial learning model. We looked across the learning domain and intended to show that a simple concept of a potential field can be easily adapted to become interesting propositions of classification and clustering models and provide further inspirations for regression and even visualisation of both classification and classifier fusion processes. The thorough analysis of the field concept led us to embody data samples into certain charge constituting the source of the field. In the case of classification, a superposition of the fields generated from each data source fully determines the potential landscape which causes testing data to fall into aroused potential wells and share the label of the samples at the bottom. We demonstrated that both attracting gravity field and electrostatic field with repelling force can be used to construct a reliable classifier, yet EFC model provides better class separation and smoother class boundaries, also reflected in the improved classification performance.

In the clustering domain we used again gravity field and an adjusted gas model based on Lennard-Jones potential. We proposed a dynamic simulation-like models of hierarchical clustering for which user can extract clusters at chosen level of granularity here determined by the time of the simulation. The gas clustering model with the controllable repelling force appears to be better in terms of class separation yet it requires an insight knowledge to properly set the parameters.

Some analogies have also been found between the field concept and the regression methodology. It has been shown that, given a small but representable training sample, using simple interpolation can quite reliably reconstruct the field, yet some problems with this approach have been identified for high-dimensional data.

Finally, the realisation that the discriminant functions used in classification are in fact certain field applied upon the data, it has been shown that both classifier as well as classifier combiner can be intuitively visualised such that it provides explanations and understanding of how the classifier or combiner arrives at the decision boundaries. Such visualisation are considered particularly vital for classifier fusion models like mean, product or majority voting combiner treated typically as very complex black boxes, now clearly visualised and transparent.

Further work into physically inspired learning models is planned to tune the growing pool of existing model prototypes presented in this paper as well as to develop new techniques harnessing analogies between energy and data evidence to manage multi-dimensional information uncertainty.

References

[1] J.A. Wheeler: Information, Physics, Quantum: The search for Links. Proc. of the Workshop on Complexity, Entropy, and the Physics of Information. Santa Fe (1989) 3-28

[2] G.J. Klir, T.A. Folger: Fuzzy Sets, Uncertainty, and Information. Prentice-Hall International Edition, (1988).

[3] W.H. Zurek: Complexity, Entropy and the Physics of Information. Proc. of the Workshop on Complexity, Entropy, and the Physics of Information. Santa Fe (1989).

[4] S. Hochreiter, M.C. Mozer: An Electric Approach to Independent Component Analysis. Proc. of the 2^{nd} International Workshop on Independent Component Analysis and Signal Separation, Helsinki (2000) 45-50.

[5] J. Principe, I. Fisher, D. Xu: Information Theoretic Learning. In S. Haykin (Ed.): Unsupervised Adaptive Filtering. New York NY (2000).

[6] K. Torkkola, W. Campbell: Mutual information in learning feature transformations. Proc. of International Conference on Machine Learning, Stanford CA (2000).

[7] K. Torkkola: Nonlinear feature transforms using maximum mutual information. Proc. of IJCNN'2001, Washington DC, USA (2001).

[8] J. Shawe-Taylor, N. Cristianini: Kernel Methods for Pattern Analysis. Cambridge University Press (2004).

[9] R.O. Duda, P.E. Hart, D.G. Stork: Pattern classification. John Wiley & Sons, New York (2001).

[10] R.P. Feynman, R.B. Leighton, M. Sands: The Feynman Lectures on Physics, Addison Wesley (1963).

[11] S.J. Cunningham, M.C. Humphrey, I.H. Witten IH: Understanding what machine learning produces part 2: Knowledge visualisation techniques (1996).

[12] T.K. Ho: Mirage - A tool for interactive pattern recognition from multimedia data. Proc. of the Astronomical Data Analysis Software & Systems XII, Baltimore, MD (2002).

[13] J. Kittler: Combininig classifiers: a theoretical framework. Pattern Analysis & Applications 1: (1998) 18-27.