Nat Comput (2009) 8:219–237
DOI 10.1007/s11047-007-9064-6

A framework for machine learning based on dynamic physical fields

Dymitr Ruta · Bogdan Gabrys

Published online: 30 October 2007
© Springer Science+Business Media B.V. 2007

Abstract Despite recent successes and advancements in artificial intelligence and machine learning, this domain remains under continuous challenge and guidance from phenomena and processes observed in natural world. Humans remain unsurpassed in their efficiency of dealing and learning from uncertain information coming in a variety of forms, whereas more and more robust learning and optimisation algorithms have their analytical engine built on the basis of some nature-inspired phenomena. Excellence of neural networks and kernel-based learning methods, an emergence of particle-, swarms-, and social behaviour-based optimisation methods are just few of many facts indicating a trend towards greater exploitation of nature inspired models and systems. This work intends to demonstrate how a simple concept of a physical field can be adopted to build a complete framework for supervised and unsupervised learning methodology. An inspiration for artificial learning has been found in the mechanics of physical fields found on both micro and macro scales. Exploiting the analogies between data and charged particles subjected to gravity, electrostatic and gas particle fields, a family of new algorithms has been developed and applied to classification, clustering and data condensation while properties of the field were further used in a unique visualisation of classification and classifier fusion models. The paper covers extensive pictorial examples and visual interpretations of the presented techniques along with some comparative testing over well-known real and artificial datasets.

Keywords Machine learning · Classification · Classifier fusion · Clustering · Data condensation · Visualisation · Gravity field · Electrostatic field · Lennard-Jones potential

D. Ruta (✉)
Intelligent Systems Research Centre, British Telecom Group, Research & Venturing, Orion Building 1st floor, pp12, Adastral Park, Martlesham Heath, Ipswich IP5 3RE, UK
e-mail: dymitr.ruta@bt.com

B. Gabrys
Computational Intelligence Research Group, School of Design, Engineering & Computing, Bournemouth University, Talbot Campus, Fern Barrow Poole BH12 5BB, UK
e-mail: bgabrys@bournemouth.ac.uk

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 (Wheeler 1989). 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 (Klir and Folger 1988). 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 (Klir and Folger 1988). The measure of fuzziness is just an example of this new dimension of uncertainty arising from the fuzzy set theory (Klir and Folger 1988). Further investigations conducted within mathematical theory of evidence revealed even more new dimensions of uncertainty (Klir and Folger 1988). 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 (Zurek 1989). The mathematical concept of a field, so commonly observed in nature, has hardly been exploited in the pattern recognition domain. Hochreiter and Mozer (2000) 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. Principe et al. (2000) 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 (Torkkola and Campbell 2000) and non-linear (Torkkola 2001) transformations of the data maximising their mutual information.

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 on both micro and macro scales. Across the artificial learning domain new field based algorithms have been developed and applied to classification, clustering and data condensation while properties of the field further reused in visualisation of classification and classifier fusion. In all of 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 and electrostatic fields we devised a new family of classification models in which data points embodied by charged particles became

the sources of the attracting or repelling field. In many variations of the model the labelled training patterns were kept either static or mobile affecting testing particles to descent down the potential gradient to ultimately meet one of the training pattern and share its label. The soft version of the model assumed particles to carry a mixture of class partitions proportional to normalised class density estimated by Parzen window method with Gaussian kernels.

Gravity field was also a starting point for new dynamic clustering models in which data-particles naturally formed hierarchical grouping along the time following their own gravitational collapse into a single cluster. Further refinement of such model has been found by reusing gas particle dynamics defined by the Lennard-Jones potential. Here the field, as for gas particles, has dual nature of attracting and repelling depending on the distance between sources. For the clustering purposes it has been made attracting on short distances and repelling on longer distances in order to ensure better cluster separation.

Based on the findings from field-based classification and clustering a new challenging problem was formulated: how to maximally condense the labelled dataset so that it maximally retains the original class densities and preserves classification performance if used for training. The problem was solved using the very same gravitational/electrostatic field-based model and showed extremely high data reduction rate under similar classification performance rates.

The field concept has also been used in the visualisation of discriminant functions of various classifiers and classifier fusion models. It advanced the classification model transparency to the level which allows to better understand and quantitatively assess the goodness of soft classification outputs and class boundaries generated by a classifier or classifier fusion system.

The paper covers extensive pictorial examples and visual interpretations of the presented techniques. A practical applicability of the models is examined and tested on 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 built upon principles of gravity field and particle dynamics in noble gases. Section 4 explains how a field can be deployed to condense labelled data without loosing the classification performance and the following section presents a methodology of classification process visualisation again with the aid of underlying field concept. In Sect. 6 some results from experiments carried out on real dataset are presented followed with the concluding remarks 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 (Shawe-Taylor and Cristianini 2004). 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 charged particles each being the source of a central field affecting other samples. All the characteristics of such a field are the results of the definition of potential and can be 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 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 model we consider a static field in a sense that the field sources are fixed to their initial positions in the input space, while the mobile testing data have cancelled out their kinetic energy at every step of the simulation. Given a training set of n data points: $X = (\mathbf{x_1}, \ldots, \mathbf{x_n})$ let each sample be the source of a field defined by a potential $V_j = -cs_i f(\overrightarrow{r_{ij}})$ 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 V_j and field interaction energy U_j in a certain point $\mathbf{y_j}$ of the input space is a superposition of the potentials coming from all the sources:

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

We can simplify this 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 Eq. 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:

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

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 the 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 in 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 as such 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 distances from the examined points and all the sources. Using matrix formulation of the problem and the mathematically oriented software like Matlab, this task can be accomplished 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 calculate 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:

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

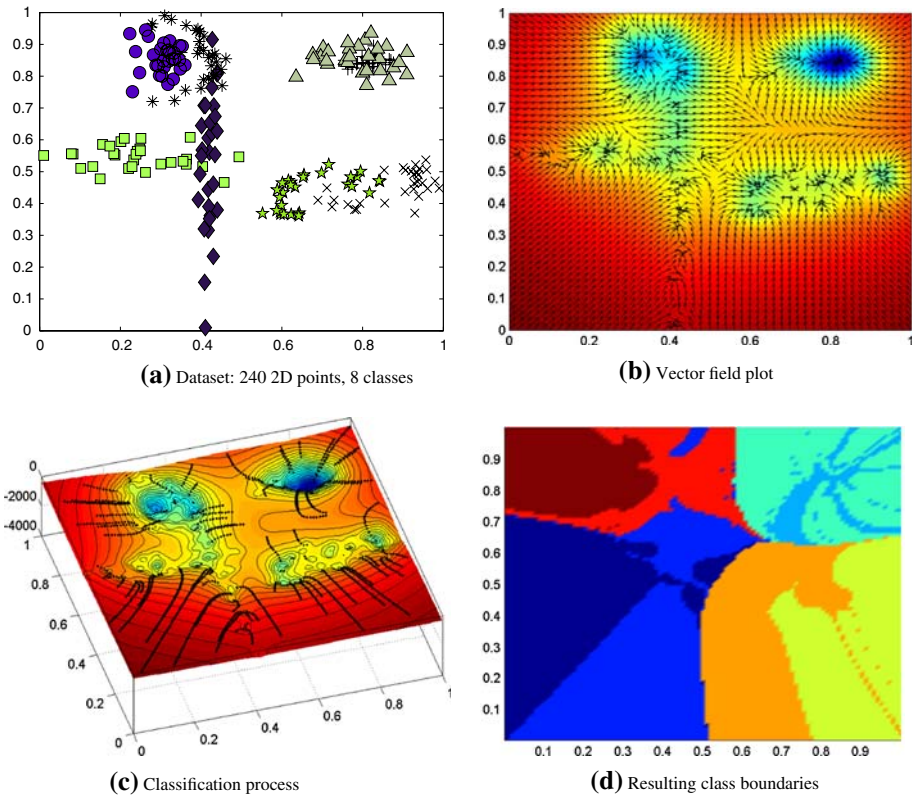
Given the distance matrix the classification process is very straightforward. The data points to be classified are simply placed in the input space and allowed to slide down the potential wells to meet one of the field source and share its label. Ignoring the dynamics of sliding data, i.e. removing the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

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

where d is an arbitrarily small step. Negative potential definition (1) ensures that testing data will always meet one of the field sources. 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-dimensional data with eight 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 the 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.



(a) Dataset: 240 2D points, 8 classes

(b) Vector field plot

(c) Classification process

(d) Resulting class boundaries

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

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

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

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 $\sum_{j=1}^{N} V_j = 0$. To achieve this goal potential definition has to be parameterised and solved with respect to the regularisation coefficient q as in the following:

$$\sum_{j=1}^{N} V_j = \sum_{j=1}^{N} \sum_{i-1}^{n} \frac{1 - q p_{jl_i}}{|\mathbf{r_{ij}}|} = 0 \tag{6}$$

In this 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 the electrostatic field classifier or shortly EFC. Example of such a field is presented in Fig. 2 again showing the whole classification process applied to the same dataset as in Fig. 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

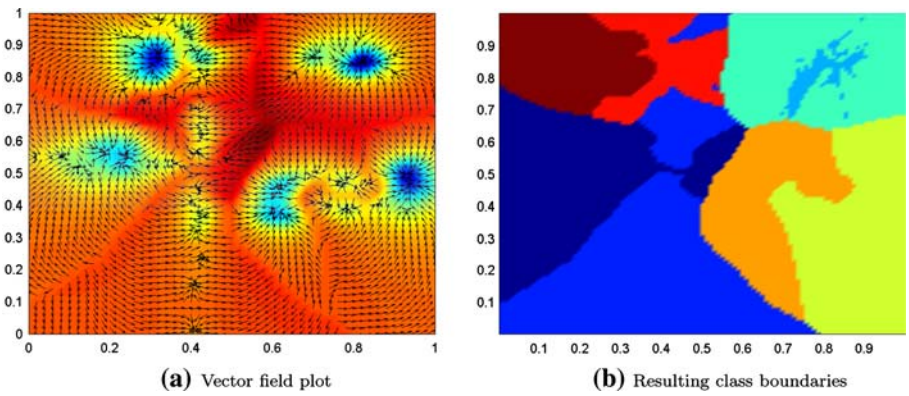


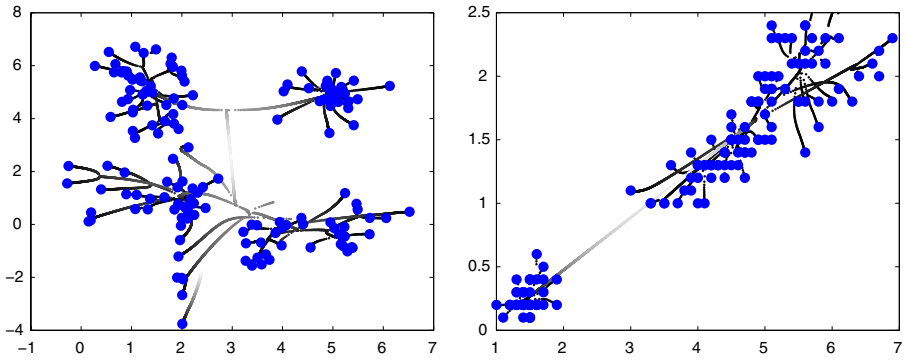
Fig. 2 Visualisation of the electrostatic field based classification process performed on the 8-classes 2D artificial data of 240 samples. (**a**) Vector plot of resulting field intensities; (**b**) Class boundaries generated by the EFC classifier

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 (Duda et al. 2001). 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 the Euclidean distance and again attracting properties of the physical fields can be reused in an attempt to construct a field based hierarchical clustering algorithms.

3.1 Hierarchical gravity based clustering model

In the simplest form we reuse the 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 continue to move according to the gravity mass dynamics equations. As before the kinetic energy that moving samples gain after each 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 which 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 Eqs. 1–4 with the only difference that the field charge multiplier s_i equal to the number of original data points in the ith 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 Fig. 1 as



(a) 8 class 2D dataset - dynamic dendrogram **(b)** Iris dataset - dynamic dendrogram

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

well as the very well known Iris dataset groups here into converging clusters at different point of time, which could correspond to the dynamic equivalent of a 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 Sect. 2.2. Here again the inspirations have been found in nature by copying gas molecules dynamics ruled by the Lennard-Jones potential (Feynman et al. 1963). In noble gasses the interaction potential between a pair of molecules is given by:

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

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:

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

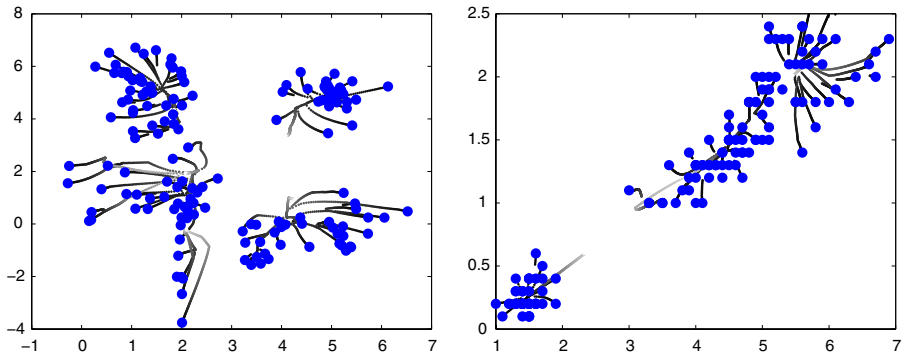
The force being identical to the field intensity vector has been simplified to the following form:

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

which retains its character yet is simpler to interpret and handle computationally for lower powers of a and b assuming $a < b$. From (8) 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 balance between repelling and attracting field can be obtained by enforcing exclusive potential definitions on different distance ranges. In our final gas model we define the following field:

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

where the field is attracting in the whole distance range excluding the range (σ_1:σ_2) where the field is repelling. The logical operators appearing in the numerators in (10) are assumed to return 1 if they are true and 0 otherwise. Figure 4 shows an example of such a model with the parameters $a = 1$ and $b = 2$. Note that for the Iris dataset the presented clustering model correctly grouped dataset into three 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.



(a) 8 class 2D dataset - dynamic dendrogram **(b)** Iris dataset - dynamic dendrogram

Fig. 4 Visualisation of the hierarchical clustering using gas model based on Lennard-Jones potential with the repelling barrier in the range of (**a**) 3–6 and (**b**) 1–3

4 Dynamic data condensation

The data fields used in the above classification and clustering models can also be used to condense the labelled data in the process of dynamic field interaction with its sources. For this purpose the field sources carrying different class charges are continually acting upon each other moving along dynamically changing field forces towards decreasing potential in the multidimensional input space. Whenever two or more data points meet, as a result of such repositioning, they instantly become a single data point with the summed charge strengthening the field around the new point yet increasing also its *mass* which reduces relative ability to shift. Like in the clustering model the whole process becomes in fact a simulation in which data-points move towards each other and gradually merge, thereby performing the act of condensation. Unlike for clustering, however, the class label charges carried by the data very strongly influence the field and its direction around the samples.

The dynamic condensation process is defined as a sequence of simulation steps which start from the original data locations and finish when no more changes in the data positions is recorded after each step or it can be stopped arbitrarily by the user at desired level of data condensation. As for the previous models the undesirable kinetic energy that would normally be gained as a result of such simulation is cancelled out after each simulation step. At each step the field and corresponding force vectors are recalculated at all existing data point locations. Using the force vectors the points are then shifted by a small step d in the directions determined by the force vectors. After the shifting phase all the data are tested for mergers which are assumed to happen when the distance between two points is lower than an arbitrarily small merger range for simplicity set to be equal to the step parameter d.

Given the distance matrix computed efficiently according to (3) the process of data evolution is very straightforward. The data points are simply sliding down along the gradients of potential they have created to meet one or more of the neighbouring field sources. The real challenge now lies in defining what happens with their original class charges during the simulation and how to resolve merging different class charges. This problem could be approached differently within the following two families of models which differ on whether the elementary class charge is allowed to be partitioned or not.

4.1 Crisp Dynamic Data Condensation (DDC_C)

Let us first assume that an individual data point is an indivisible fault free unit of evidence. During the condensation process these data points are let free and as they move and merge the field changes according to their new locations. Depending on how the class labels of the data are used the crisp electrostatic condensation models can be further subdivided into unlabelled and labelled which differ fundamentally from each other.

Let us first consider Crisp Unlabelled DDC (DDC_{CU}) model which uses attracting-only data field as shown in (1). To be able to effectively use such a field for the condensation purposes the dynamic simulation model has to be applied to all classes in isolation as otherwise all the data merge to the centre-mass point and completely destroy the class structure. The simulation can be carried out simultaneously for all the classes yet in that case the intra-class interactions have to be insensitive to the other classes interactions. In this model data points of each class are collapsing gradually up to a single data point per class which terminates with the weights proportional to the counters of original points in the condensed data.

In the Crisp Labelled Electrostatic Condensation (DDC_{CL}) the labels of the training data are used to determine the sign of the potential. Negative potential is generated by the data from the same class and positive for the data from different class such that the data points from the same class are attracting each other and data from different classes repel from each other. Note that during the condensation process all the labels of the condensing dataset are known hence the definition of potential becomes a special case of (5) where all the charge partitions p_{ji} are crisp i.e. are either 1 or 0. An advantage of this approach for classification purposes is that during such process the data from different classes would try to separate from each other. However, in case of elongated class shapes some data tails or other isolated data points could be pushed away by the dense neighbouring regions from different classes. Another problem that emerges in the labelled electrostatic model is collisions of data from different classes. Such collisions could happen if the attracting forces generated by large data concentrations from the same class override weaker repelling actions from the samples of different classes. Such collisions could be resolved using Parzen class density as a preference in choosing the winning class.

4.2 Soft Dynamic Data Condensation (DDC_S)

It is reasonable to assume that the labelled classification data is noisy and often faulty and hence to treat it in the soft probabilistic terms. The data labels are summarised in the form of soft class membership or partition values. These class partitions can be obtained from the original class densities using Parzen window density estimator and then mapped onto the probabilities. According to Parzen-window approach (Duda et al. 2001) an estimate of the data density in point x_j can be obtained by calculating:

$$p(\mathbf{x_j}) = \mathbf{p_j} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\frac{\mathbf{x_j} - \mathbf{x_i}}{h_n}\right) \tag{11}$$

where V_n is a window volume, φ is a specific window function and h_n is a smoothing parameter. This model uses the common Gaussian window function defined as follows:

$$\varphi(u) = \frac{1}{2\pi} e^{-u^2/2} \tag{12}$$

The leave-one-out maximum likelihood estimation is applied to find the optimal smoothing parameter h_n. Let p_k denote a vector of Parzen density estimates measured on dataset X but generated only from kth class of the dataset X, $k \in (1,\ldots,C)$. Class partitions are obtained from Parzen densities using typical transformation mapping used in classification and scaled up to sum up to a unit:

$$p_k^N = \frac{1}{1+e^{-p_k}}, \qquad p_k^S = \frac{p_k^N}{\sum_{i=1}^{C} p_i^N} \tag{13}$$

Given the matrix of partitions as defined above, one could easily apply electrostatic potential (5) yet there is still a freedom in deciding how to control the variability of such soft labels during the condensation process. The most conservative model called Soft Fixed-Field Condensation (DDC_{SFF}) assumes a fixed field built on the original data which is kept unchanged during the condensation process. Such approach would try to maximally preserve the original Parzen density structure even if it requires continuous relabelling of moving samples. What happens is that as data-points move, their class charge partitions dynamically change depending on the new locations at which new class densities are calculated. Here the merging process is free of any conflicts as the class partitions of the merging samples are simply adding up to the new "heavier" sample which stays partitioned after the merger.

The Soft Fixed-Labels Condensation (DDC_{SFL}) keeps label partitions fixed during the condensation, whereas the field is allowed to change freely depending on actual data locations. The merging process remains additive with respect to class charges as in the previous case.

Finally the least constrained Soft Dynamic Data Condensation (DDC_S) model releases all the constraints and lets both the field and class partition charges change as the data samples evolve during condensation process. As in previous soft models colliding data merge into a single point with summed class label partitions.

For all soft condensation models the field is continuously re-normalised using a regularisation parameter calculated as shown in (6). Both soft and crisp condensation processes finish when the sum of data shifts at a single step is smaller than an arbitrarily small distance, for simplicity chosen to be equal to d used in (4).

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 decision making 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 remember. Surprisingly there is not much evidence of this fact being taken advantage of in commercial applications of pattern recognition. Many visualisation techniques have been proposed in the sister domain of knowledge discovery (Cunningham et al. 1996), yet even huge analytical packages like Mirage (Ho 2002) miss out on 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 $\Omega = \{\omega_1,\ldots,\omega_C\}$, 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 each class. Examples of such discriminant function plots are shown for a decision tree and a quadratic classifier in Fig. 5. The final classification decision is formulated as:

$$\omega_d = \arg \max_{j=1}^{C} P(\omega_j \mid \mathbf{x}) \tag{14}$$

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

$$D = \max_{j=1}^{C} P(\omega_j \mid \mathbf{x}) \tag{15}$$

Decision surface can be easily picked up from Fig. 5, 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 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 ith (i = 1,…,N) classifier applied to the class ω_j (j = 1,…,C). The major classifier fusion operators (Kittler 1998) can easily be expressed by simple aggregation operations

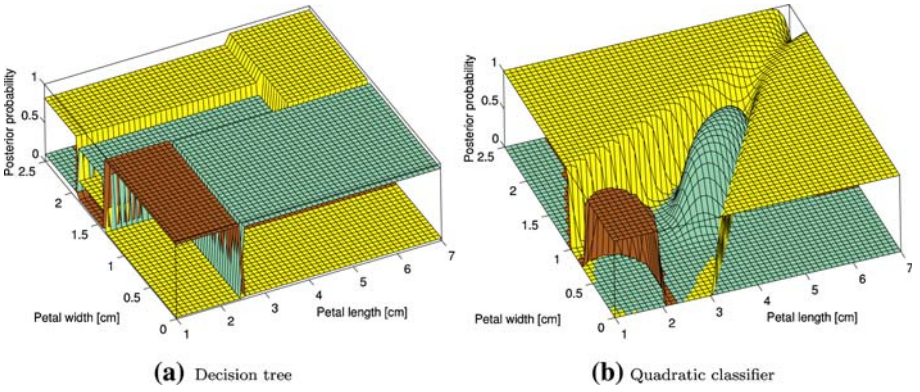


Fig. 5 Visualisation of the discriminant functions applied to the 3-class Iris dataset. (**a**) Discriminant functions generated by the decision tree classifier; (**b**) Quadratic classifier discriminant functions

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:

$$D_F = \max_{j=1}^{C} F_{i=1}^{N} P_{ij} \tag{16}$$

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 6a shows an example of the complete visualisation of the decision tree classifier, while Fig. 6b–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

The presented physical field based learning models have been evaluated in the context of classification and condensation problems. Both GFC and EFC classifiers have been tested over six datasets along with 15 other classifiers for comparison. Figure 7 shows the details of datasets and classifiers used in this experiment. For the first four 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 is particularly the case of EFC classifier that introduces inter-class repelling action and soft partitioning of class charges. This attribute makes it an interesting alternative to other classifiers that could be particularly useful in combining classifiers.

The strength of condensation capability of the presented family of condensation models has been evaluated in terms of classification performance obtained at different levels of data reduction and then compared with the performance on the original data. All the models have been applied to the well-known Land Satellite Image dataset from UCI Repository.[1] The training set consisted of 2,000 data points and the remaining 4,435 data

[1] University of California Repository of Machine Learning Databases and Domain Theories, available free at: ftp.ics.uci.edu/pub/machine-learning-databases

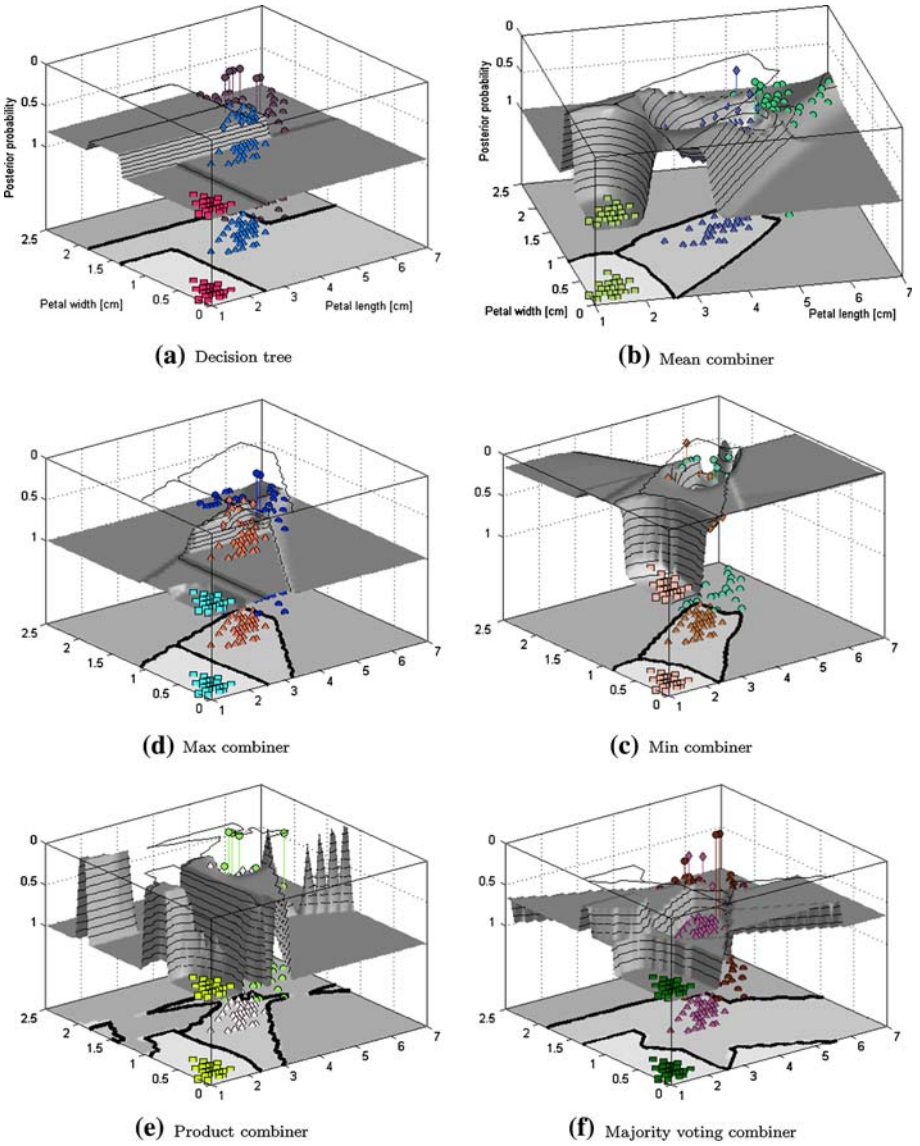


Fig. 6 Complete visualisation of the decision tree classifier and mean combiner built on decision tree, quadratic and neural net classifiers applied to iris dataset

points were used as a testing set. All the dynamic condensation models have been applied to the multiple training sets collected during the condensation at increasing level of dataset reduction. These training sets were used to train the two selected benchmark classifiers: Parzen density classifier (PDC) and k nearest neighbour (KNN). The trained classifiers were then tested on the testing set to produce the performance condensation profiles as shown in Fig. 8.

The soft condensation methods clearly outperformed the crisp methods with the constrained versions of DDC_{SFF} and DDC_{SFL} consistently taking the leading positions. In the

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

Fig. 7 A list of 6 datasets (**a**) and 15 classifiers (**b**) used in experiments for performance comparison with GFC and EFC classifiers

following experiment the best dynamic condensation method i.e. DDC_{SFF} was compared with the typical random sampling, condensation based on k-means intra-class clustering, and the state-of-the-art RSDE density estimation method described in (Girolami and He 2003). Both k-means and RSDE have been applied to reduce each class separately yet for RSDE it was only possible to obtain a single point of the condensation profile i.e. for the condensation level at which the RSDE method converged. For k-means clustering, the condensation level was controlled by selecting the numbers of cluster centres for each class that were proportional to the prior class probability. The random sampling was applied to the incrementally reduced training sets preserving the prior class probabilities. The results of this experiment are shown in Fig. 9.

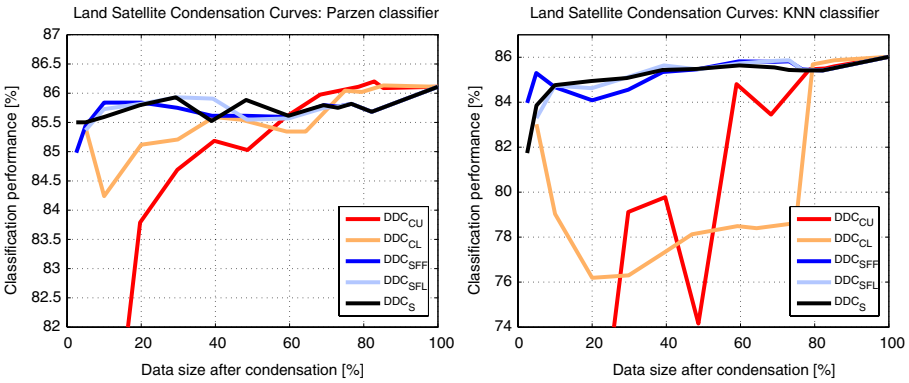
Clearly DDC_{SFF} outperformed the other methods and even showed higher testing performance than the model trained on the data reduced by the top-in-the-field RSDE method (Girolami and He 2003). The top-ranked condensation algorithm managed to retain close to the 99% of the original classification performance for almost 99% reduced training set. An interesting observation can be made on the basis of the reduced set visualised in Fig. 9 which shows that some classes were partitioned into isolated clusters as a result of the repelling inter-class forces. The class remainders retained their identity outside of the main class mass yet they found the positions which do not disturb the purity of the main class densities as opposed to the fixed-position condensation methods which would most likely ignore the class remainders treating them as outliers. This property could be the key to retaining maximum class representativeness at high data condensation levels.

7 Conclusions

In this article, 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 form interesting propositions for classification, clustering and data condensation models and provide further inspirations for visualisation of both classification and classifier fusion processes. The thorough analysis of the field concept led us to treat data samples as particles carrying certain charge and acting as sources of the class-related field.

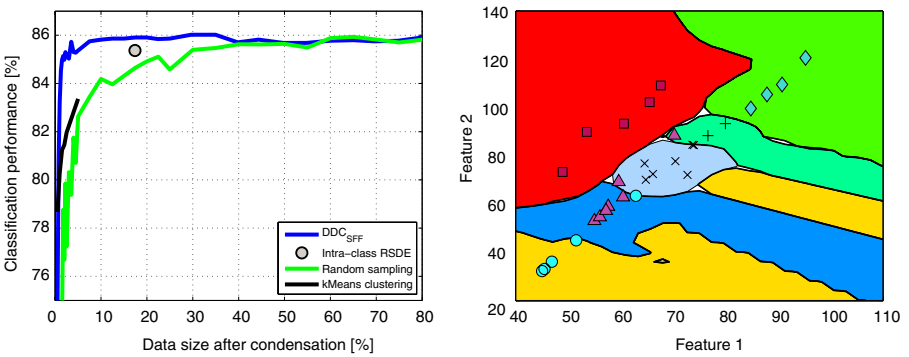
Table 1 10-fold cross-validation error rates (%) obtained for 17 classifiers and 6 datasets described in Table iments

	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



(a) Parzen density classifier **(b)** K nearest neighbour classifier

Fig. 8 Visualisation of the classification performance obtained on increasingly condensed Land Satellite datasets at different condensation levels for (**a**) Parzen density classifier; (**b**) K nearest neighbour classifier



(a) Condensation profiles comparison **(b)** Landsat data after 99% condensation using DDC_{SFF}

Fig. 9 Visualisation of the Parzen density classifier performance obtained on increasingly condensed Land Satellite datasets using DDC_{SFF}, intra-class RSDE method, random sampling and k-means clustering. (**a**) Condensation profiles comparison; (**b**) Visualisation of the LandSat dataset reduced by 99% using DDC_{SFF}, overlayed on the class boundary map of a Parzen density-based classifier trained on the full training set

In 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 source samples at the bottom of a well. The electrostatic metaphor of charge-dependent attracting or repelling field was adopted to generate intra-class attraction and inter-class repulsion to further encourage class separability and smooth 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 from which the user can extract clusters at a 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 scaling the range of attracting and repelling interaction.

We showed that the field concept has also found a fruitful application in data condensation. The proposed family of soft dynamic condensation models allows for mergers of data from different classes yet accumulatively retaining soft/fuzzy class partitions. In the most successful DDC_{SFF} condensation model, the field is directly guided by the Parzen estimates of the original class densities which continuously redistribute class partitions as the data merge and move during the condensation process. The new approach of dynamic condensation gives an option to shift the complexity of labelled data from the input space to the class partition space and it has been shown that for classification purposes that large sets of crisp-class data are as good as a very small, sometimes even 99% reduced, data with in-build soft class partition representation.

Finally, the realisation that the discriminant functions used in classification are in fact a certain field acting 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 and soft classification outputs. Such visualisation is particularly vital for classifier fusion models typically considered 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 further analogies between physical and natural phenomena and the analytics of the machine learning.

References

Wheeler JA (1989) Information, physics, quantum: the search for links. Proc of the Workshop on Complexity, Entropy, and the Physics of Information. Santa Fe 3–28

Klir GJ, Folger TA (1988) Fuzzy sets, uncertainty, and information. Prentice-Hall International Edition

Zurek WH (1989) Complexity, entropy and the physics of information. Proc of the Workshop on Complexity, Entropy, and the Physics of Information. Santa Fe

Hochreiter S, Mozer MC (2000) An electric approach to independent component analysis. Proc of the 2nd International Workshop on Independent Component Analysis and Signal Separation, Helsinki 45–50.

Principe J, Fisher I, Xu D (2000) Information theoretic learning. In: Haykin S (ed) Unsupervised adaptive filtering. New York

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

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

Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press

Duda RO, Hart PE, Stork DG (2001) Pattern classification. John Wiley & Sons, New York

Feynman RP, Leighton RB, Sands M (1963) The Feynman lectures on physics. Addison Wesley

Cunningham SJ, Humphrey MC, Witten IH (1996) Understanding what machine learning produces part 2: Knowledge visualisation techniques

Ho TK (2002) Mirage—A tool for interactive pattern recognition from multimedia data. Proc of the Astronomical Data Analysis Software & Systems XII, Baltimore, MD

Kittler J (1998) Combining classifiers: a theoretical framework. Pattern Anal Appl 1:18–27

Girolami M, He C (2003) Probability density estimation from optimally condensed data samples. IEEE Trans Pattern Anal Machine Intell 25(10):1253–1264