

Physical field models for pattern classification

D. Ruta, B. Gabrys

126

Abstract Recent findings in pattern recognition show that dramatic improvement of the recognition rate can be obtained by application of fusion systems utilizing many different and diverse classifiers for the same task. Apart from a good individual performance of individual classifiers the most important factor is the useful diversity they exhibit. In this work we present an example of a novel, well performing non-parametric classifier design, which shows a substantial level of diversity with respect to other commonly used classifiers. Inspired by the mechanics of omnipresent physical fields like gravitational or electrostatic, we considered the data as particles carrying elementary units of charge. The charge has been presented as a source of the potential triggering attracting interaction among the data. This interaction has been reformulated as data matching procedure and developed into original classification technique where the unlabelled testing data are captured by the labelled training data and share their labels. In the extended model apart from the spatial data distribution we also exploit topology of class labels to devise repelling force as field action between differently labelled data samples. As we show introduction of the repelling force clearly smoothes the decision boundaries and improves performance while still preserving attractive diversity properties of the classification model. The paper covers extensive examples and visual interpretations of the presented techniques supported by the experimental work with established datasets and classifiers.

1

Introduction

Research in pattern recognition proves that no individual method can be shown to be the best for all classification problems [1, 2, 11]. Instead, an interesting alternative is to construct a number of diverse, generally well performing classification methods, and combine them on different levels of classification process. Combining classifiers has been shown to offer a significant classification improvement for some non-trivial pattern classification problems [1, 11]. However, the highest

improvement of a multiple classifier system is subject to the diversity exhibited among the component classifiers [5, 7, 10]. In this paper we propose an example of such a novel classifier that performs well individually and can be shown diverse with respect to other commonly used classifiers. In designing the new classifier we exploit a notion of a static field generated by a set of samples treated as physical particles. Our approach is closely related to the idea of an *information field*, recently emerging from the studies in information theory, where increasingly deep analogies are drawn with the physical world [4, 14]. Shannon entropy representing probabilistic interpretation of information content is an example of a direct counterpart to the thermodynamic entropy related to physical particles. Information or its uncertainty is quite often compared to energy, with all its aspects [4]. The latest findings led even to the formulation of the quantum information theory based on well-developed quantum physics [14]. The mathematical concept of a field so commonly observed in nature, has hardly been exploited for pattern recognition purposes. In [3], Hochreiter and Mozer apply the electric field metaphor to the independent component analysis (ICA) problem where joint and factorial density estimates are treated as a distribution of positive and negative charges. In [6], Principe et al. introduce the concept of information potentials and forces employing unconventional definition of mutual information based on Renyi's entropy. Torkkola [12] further used these concepts for linear [12] and non-linear [13] transformations of the data maximising their mutual information.

Inspired by these tendencies we directly adapt the concept of the field to represent the labelled data, and develop it further to be potentially applied for classifier fusion. The foundation of this idea is the fact that quite often non-parametric classification techniques based on clustering or matching prove to be very successful for a wide variety of problems without prior knowledge about the data. However, the nearest neighbour (NN) approach leads commonly to data overfitting and in k-NN the influence of the distance variability is ignored. In another non-parametric example of Parzen density based classification [2] these drawbacks are partially resolved but still the winning class density effectively carries information from only a limited number of training samples coming from this class. Moreover, inclusion of all the labels from the training data, regardless of the spatial data distribution, may in some situations lead to the propagation of noise harming the overall performance.

Published online: 14 July 2003

D. Ruta (✉), B. Gabrys
Applied Computational Intelligence Research Unit,
University of Paisley, High Street,
Paisley PA1-2BE, Scotland, UK
e-mail: ruta-ci0@paisley.ac.uk

To ensure that every data sample is actively contributing in the formation of the final classification decision the training data can be considered as charged particles each being a source of a certain field. All the characteristics of such field are the results of the definition of potential and can be absolutely or 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 demand seems to be satisfiable within the data field framework remembering that the gradient of the field potential results in the force which could move the mobile testing sample and label it by sharing the label of any fixed training sample met on its way. 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. 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.

Note that in this fashion the information about the labels do not affect the trajectories of the testing data and in this sense the mechanism is purely unsupervised. To incorporate the information about the labels of the training data and still retain the data driven field model the repelling force can be introduced by direct analogy to the electrostatic field. In such a design a pair of samples from different classes should produce a positive potential, giving a repelling force, whereas samples from the same class should attract each other as a result of a negative potential. To incorporate both data characteristics of the unlabeled set and label information, a purely data driven attracting field should be updated by the opposing intra-class attracting field and inter-class repelling field, the balance of which can be estimated by the Parzen data density method.

The remainder of the paper is organized as follows. Section 2 explains the way in which the data is used as sources of the attracting field, including implementation details of the gravity field classifier. The next section considers the extension of the model into electrostatic design allowing for both attracting and repelling force. In Sect. 4 we explain the mining of diversity including its role for combining classifiers and simple ways of measuring it. Section 5 presents the comparative experimental results carried out with real datasets. Conclusions and plans for future work are briefly presented in the closing section.

2

Gravity model of the data attracting field

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 samples: $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, let each sample be the source of a field defined by a potential

$$U_j = -cs_i f(\vec{r}_{ij}) \quad (1)$$

Where c represents the field constant, s_i stands for a source charge of the data point \mathbf{x}_i , and $f(\vec{r}_{ij})$ is a certain non-negative function decreasing with an increasing length of the vector \vec{r}_{ij} describing the distance between the source \mathbf{x}_i and the point \mathbf{y}_j in the input space. Note that the potential is always negative, which indicates the attracting properties of the data. In this section we adopt the gravitational field for which:

$$f(\vec{r}_{ij}) = \frac{1}{|\vec{r}_{ij}|} \quad (2)$$

For notation simplicity we assume: $r_{ij} = |\vec{r}_{ij}|$. The overall potential in a certain point \mathbf{y}_j of the input space is a superposition of potentials coming from all the sources:

$$U_j = -c \sum_{i=1}^n \frac{s_i}{r_{ij}} \quad (3)$$

Considering a new j th data point in such a field, we can immediately associate with it energy equal to:

$$E_j = -cs_j \sum_{i=1}^n \frac{s_i}{r_{ij}} \quad (4)$$

Simplifying the model further we can assume that all data points are equally important and have the same source charge equal to the unit: $s_i = 1$, thus eliminating source charges from equations (3) and (4). Another crucial field property is its intensity, which is simply a gradient of the potential and can be formally expressed by:

$$\mathbf{E}_j = -\vec{\nabla} U_j = -\left(\frac{\partial U_j}{\partial y_{j1}}, \frac{\partial U_j}{\partial y_{j2}}, \dots, \frac{\partial U_j}{\partial y_{jm}} \right) \quad (5)$$

Solving Eq. (5) leads to the following form of the field vector:

$$\begin{aligned} \mathbf{E}_j &= -c \left(\sum_{i=1}^n \frac{y_{j1} - x_{i1}}{r_{ij}^3}, \dots, \sum_{i=1}^n \frac{y_{jm} - x_{im}}{r_{ij}^3} \right) \\ &= -c \sum_{i=1}^n \frac{\mathbf{y}_j - \mathbf{x}_i}{r_{ij}^3} \end{aligned} \quad (6)$$

A vector of field intensity (simply *field vector*) shows the direction and the magnitude of maximum fall of the field potential. By further analogy to the gravitational field, the charged data point is affected by the field in the form of a force attempting to move the sample toward the lowest

energy levels. As the charge has been assumed uniform and of unit value and so excluded from the equations, the force vector and intensity (field) vector become the same:

$$\mathbf{F}_j = \mathbf{E}_j \quad (7)$$

The only difference between them is a physical unit, which is of no importance in our application. The concept of field forces will be directly exploited for the classification process described in Sect. 2.3. The field constant c does not affect the direction of forces but only determines their magnitudes. As previously, without any loss of generality we can assume its value as unitary and in that way free all the field equations from any parameters, apart from the definition of the distance itself.

2.1 Field generation

From a perspective of classification, the generation of a field could represent the training process of a classifier. However, as the training data uniquely determines the field and all its properties, the training process may be omitted or just reduced to acknowledging the training data as field sources. All the calculations required to classify new data are carried out online during the classification process avoiding any imprecision caused by approximations that might have been applied otherwise. It is very similar to the generation and operation of the very well known k -nearest neighbor classifier. In case of a large amount of data to be classified, the data-grid option is available although not completely precise as in the previous case. Namely, one can split the input space into small hyper boxes and calculate all the field properties required in the center of each hyper box. In this way the field can be approximated at any point just by local aggregation procedures. The training process would be substantially prolonged, but the classification phase would require calculations related to just one or two points from the neighborhood and therefore can be drastically shortened. For both methods, the critical factor is calculation of the distances from the examined points and all the sources. Using a matrix formulation of the problem and the appropriate mathematical software, this task can be completed rapidly even for thousands of sources.

Let $Y^{[N \times m]}$ denote the matrix of N m -dimensional data points at which we want to examine the field, and $X^{[n \times m]}$ be the matrix of n training data field sources. The task is to obtain the matrix $D^{[N \times n]}$ of all the distances between the examined data and the training data. As opposed to the time consuming double loop implementation, introducing a matrix formulation leads to significant savings in terms of code length and processing time. That is, matrix D can be calculated thus:

$$D = Y \circ Y \bullet \mathbf{1}^{[1 \times n]} - 2 \bullet Y \bullet X^T + \mathbf{1}^{[N \times 1]} \bullet X \circ X \quad (8)$$

where “ \circ ” denotes the operator of element-wise matrix multiplication (multiplication of corresponding elements), “ \bullet ” represents standard matrix multiplication, and $\mathbf{1}^{[n \times m]}$ stands for a matrix of size $n \times m$ with all elements equal to one. Implementation of the above rule under Matlab 5 is 20 times shorter to calculate than the double loop algorithm. As an example, it took a P4 processor 8 seconds to

calculate all distances between 1000 1000-D points and another set of 1000 1000-D sources. Given the distance matrix D all the properties of the field can be obtained by simple algebra performed on D . Avoiding numerical problems, the distances have been limited from the bottom by an arbitrary threshold d preventing division by zero.

2.2 Numerical example

As an example we generated 50 random 2-D points from a range (0,1) in both dimensions. Figure 1 provides the visualization of the field arising from the training data. Not surprisingly, the potential is the lowest in the maximum data concentrations and generally in the middle of the regions occupied by the data. This phenomenon nicely correlates with the classification objective, as the highly concentrated regions should have a greater chance of data interception. The same applies to the dramatic local decrease of the potential around the field sources. Presence of the field can also be interpreted as a specific curvature of the input space imposed by the presence of field

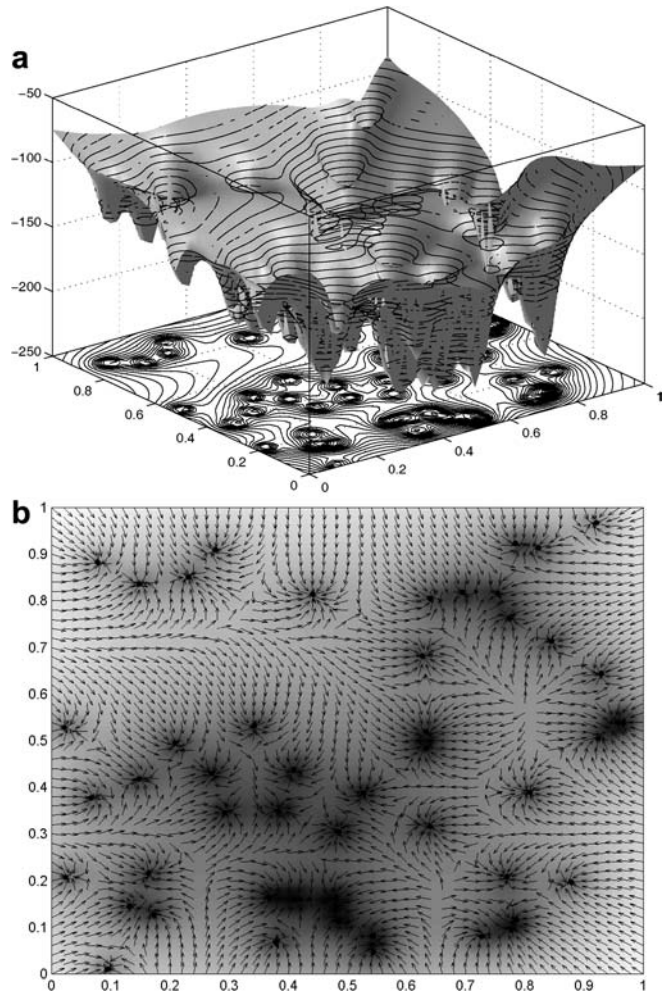


Fig. 1a, b. Visualization of the field generated by 50 2-D data drawn at random from the range (0,1). a 3-D plot of the potential, b Vector plot of the field pseudo-intensity (‘pseudo’ as the vectors point only in the true directions of the field but their lengths are here fixed for visualization clarity)

sources, that is, the data. Each point in such a curved input space will be forced to move along the force vectors (Fig. 1b) ultimately ending up in a position of one of the field sources. In this way the field built upon the data has the ability to uniquely transform the input space. For classification purposes, such a transformation leads to the split of the whole input space into subspaces labelled according to the labels of field sources intercepting the data from these subspaces. Figure 2 visualises the same field with an additional 3rd dimension of the data. It has to be noted that physical fields are restricted to only 3 dimensions of the natural reality perceived by our senses. In pattern recognition problems data could be and very often is multidimensional with tens or even hundreds of features. Dimensionality of the data does not, however, affect the field concept which mathematically can operate on any finite number of dimensions.

2.3

Classification process

Given the data field, the classification process is very straightforward and simply reduces to finding the gradient

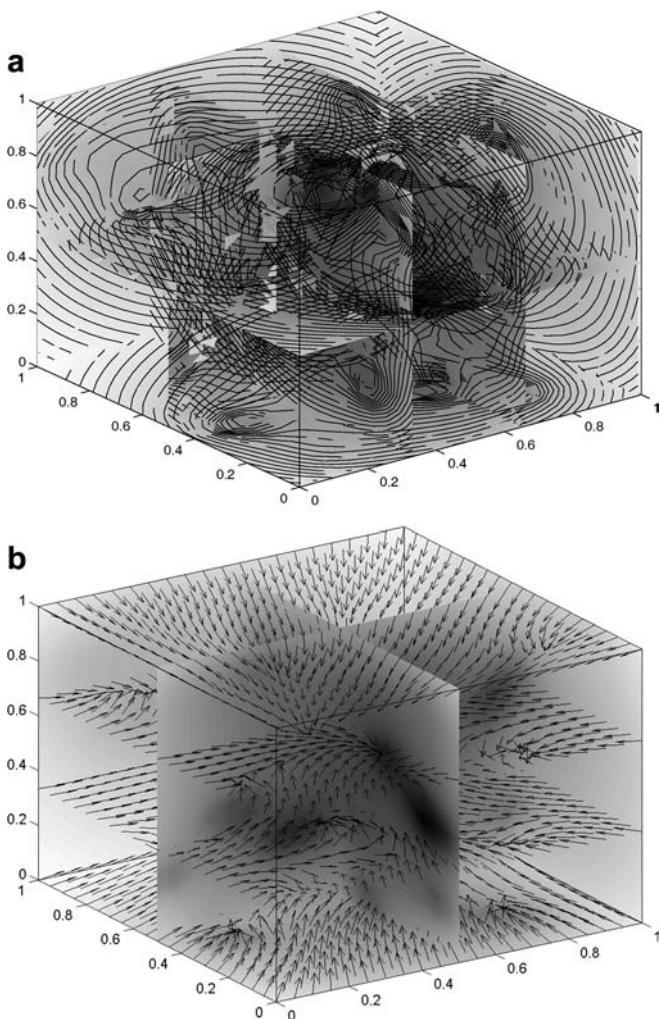


Fig. 2a, b. Visualization of the field generated by 50 3-D data drawn at random from the range (0,1). **a** Volumetric slice plot of the field potential, **b** Pseudo-vector plot on the volumetric slices illustrating true field intensity directions in the 3-D space

descent source. The slide begins from the position of a new data point to be classified. For algorithm stability we ignore the actual values of the field or force vectors, following just their local directions. The sample to be tested is shifted along this direction by a small step d and the whole field is recalculated again in the new position. This procedure is repeated until the unlabelled sample approaches any source at the distance lower or equal than d . If this happens the sample is labelled according to the source it was intercepted by. From now on we will refer to such a classification model as a gravity field classifier or GFC for short.

Although the method looks simple there are some numerical problems involved. The fixed parameter d , corresponding to the length of the shift vector, could cause different relative shifts for differently scaled dimensions. To avoid this problem we normalize the input space to cover all the training data within the range (0,1) in all dimensions and set the step d to the fixed value: d . During the classification process the new data is transformed to the normalized input space and even if its position falls outside the range (0,1) in any dimension, the step d remains still well scaled. The step d has been deliberately denoted by the same symbol used for the lower limit of the distance introduced in the previous section. This ensures that the sample never misses the source on its trajectory and additionally two parameters are reduced to just one. To speed up the classification process the step can be extended as long as the data size is small enough and the distances between the sources remain larger than d .

2.4

Matrix implementation

Rather than classifying samples one by one we used Matlab's capabilities to classify samples simultaneously. Given a distance matrix D calculated efficiently according to (8), the matrix of forces $\mathbf{F}^{[N \times m]}$ can be immediately obtained by (6). Exploiting the triangular relation between forces and shifts and given constant d , the matrix of shifts $\Delta Y^{[N \times m]}$ can be calculated by the formula:

$$\Delta Y = \frac{d \bullet \mathbf{F}}{(\mathbf{F} \circ \mathbf{F}) \bullet \mathbf{1}^{[m \times 1]} \bullet \mathbf{1}^{[1 \times m]}} \quad (9)$$

The full algorithm of the gravity field classifier can be expressed in the following sequence of steps:

1. Given labelled training data X and unlabelled testing data Y to be classified, calculate the matrix of distances D according to (8).
2. Calculate the matrix of field forces at the positions of unlabelled data to be classified.
3. Given a fixed step, calculate the shifts of the samples according to (9).
4. Shift the samples to the new locations calculated in the step 3.
5. For all samples check if the distance to any source is less or equal to the step d . If yes, classify these samples with the same labels as the sources they were intercepted by and remove them from the matrix Y .
6. If matrix Y is empty finish else go to step 1.

Transformation as presented above leads to the split of the whole input space into subspaces labelled according to the labels of field sources capturing the data from these subspaces. Figure 3 presents a graphical interpretation of the classification process for an artificial dataset with 8 classes. One can notice that the information about the labels of the training data is not used until the very end of the classification process. This property makes the method an interesting candidate for an unsupervised clustering technique. The class boundary diagram reveals an interesting effect of the presented classification method. Namely, occasionally one can observe a narrow strip of one class getting deep into the area of another class. This is the case of the *potential ridge*, which is balanced from both sides by the data causing the field vector to go in-between, sometimes even reaching another class. Although this phenomenon is not particularly desirable for an individual classifier and may be harming its performance, as we show in the experiments it rarely occurs and contributes to the satisfactory level of the diversity the SFC classifier exhibits with other classifiers.

2.5

Comparison with other classifiers

The static field classification presented in this paper shares some similarities with other non-parametric classifier designs. The process of field generation can be seen in fact as an indirect parametric estimation of the data density where kernels are defined by potentials generated by each training data sample. Although technically similar, the two approaches are quite distinct in the use of the calculated quantities. If applied to the labelled dataset, the data density is calculated separately for each class and further compared with each other to decide the winning class. This strategy is also closely related to Bayesian classifiers, which pick a class with maximum a posteriori probability calculated on the basis of assumed or estimated class probability density functions. In the case of static field classification, the data potential is calculated using all the training samples regardless of their labels. The curved geometry of the input space forces testing samples to fall down the local potential well to meet one of the sources and share its label. Such an approach also resembles

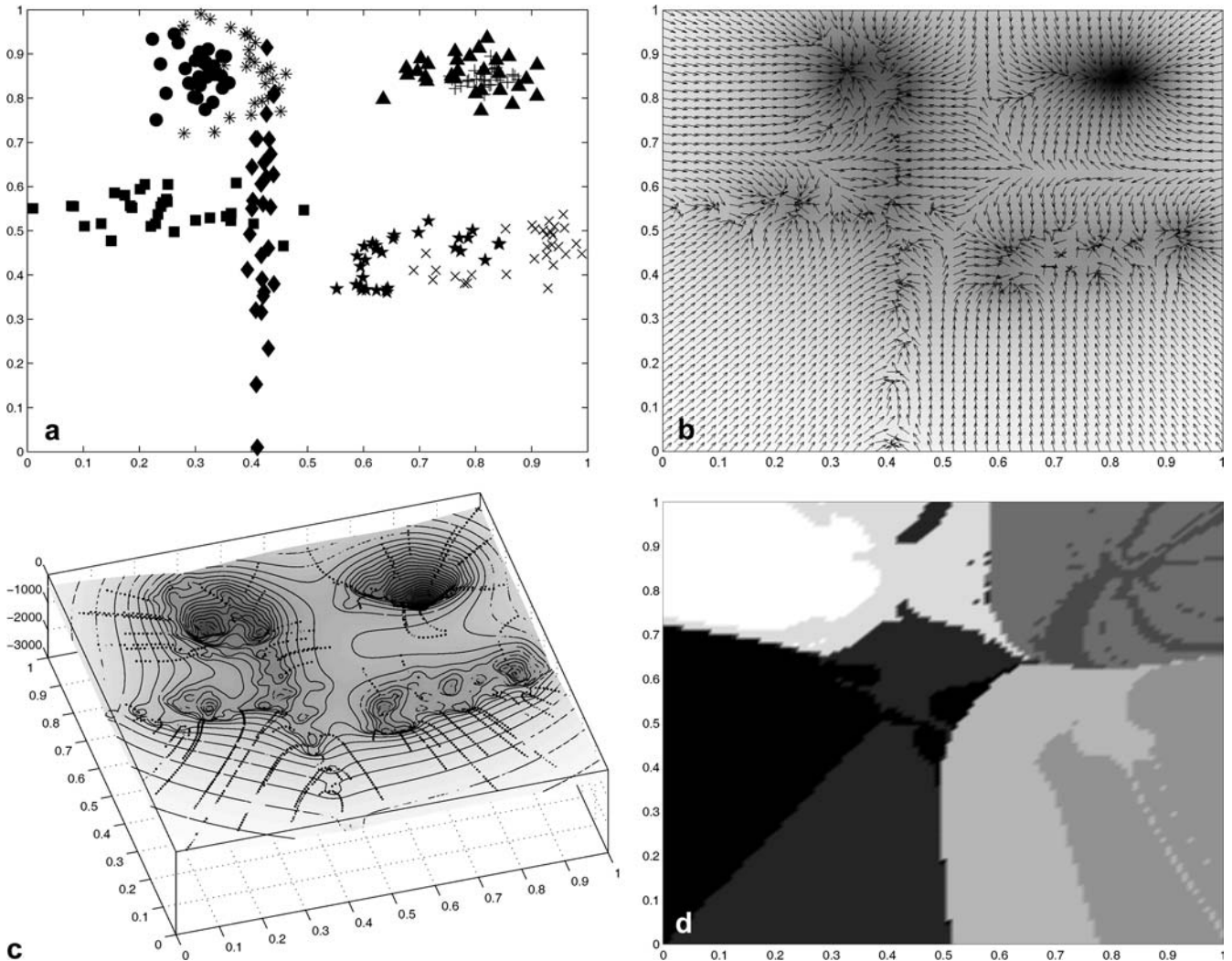


Fig. 3a–d. Visualization of the static field based classification process performed on the 8-class 2-D artificial data of 240 samples. a Scatter plot of the training data, b Vector plot of the field

pseudo-intensity, c Trajectories of exemplary testing data sliding down along potential gradients, d Class boundaries diagram

unsupervised clustering techniques, where spatial data distribution is the only information used for data matching [2]. Up to the labelling process, the GFC classifier remains purely unsupervised and for that reason is a good candidate as an original clustering technique.

The classification process of falling into a potential well of a single source strongly resembles the nearest neighbour rule, where the label is passed from the closest training samples found in the input space. In fact if the field potential is defined to be very strong locally, the decision boundaries from such a classifier should converge to the Voronoi diagram [2] reflecting the nearest neighbour rule. The field approach appears to be more general, flexible and addressing the complete landscape of the spatial data distributions rather than just the local region.

Figure 4 provides a comparison of the class boundaries generated by the static field classifier, the Parzen density based classifier and the nearest neighbour technique.

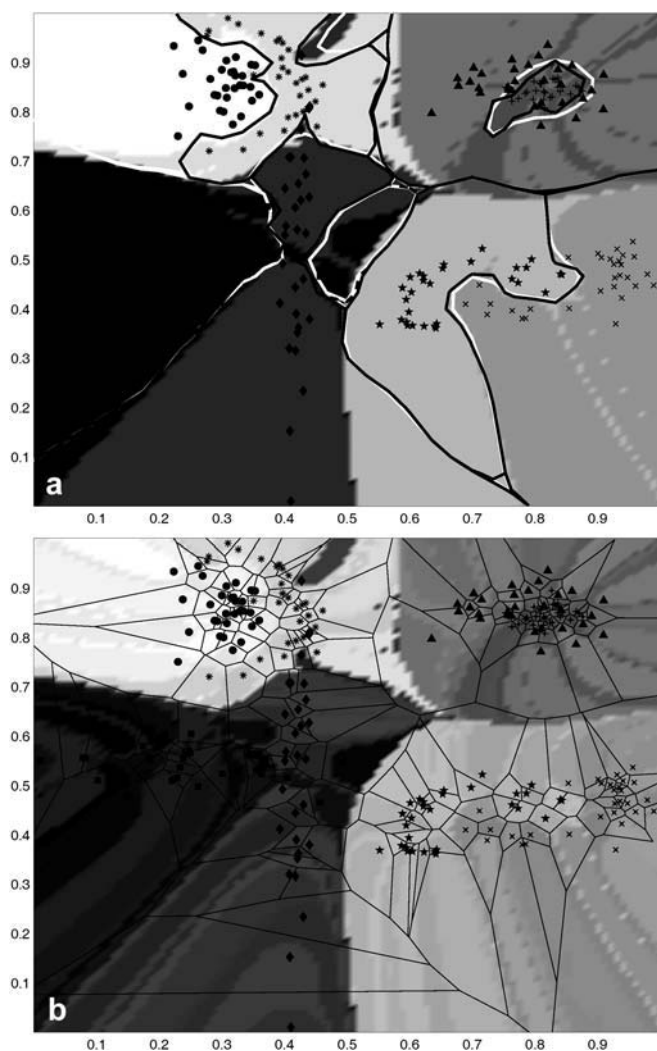


Fig. 4a, b. Comparison of class boundaries of non-parametric classifiers. **a** Decision boundaries of GFC (coloured filled regions) compared to kNN (white solid lines) and Parzen classifier (black solid lines) with data scatter plot, **b** Voronoi diagram (NN rule) compared with equivalent single sample interception regions for GFC shown in different shades within the corresponding classes

Given artificial dataset of 240 2-D samples with 8 classes, the boundaries clearly form similar shapes but have quite different exact boundary lines. What is surprising is an almost exact match of the boundaries from the Parzen and NN classifiers. Much bigger differences can be seen when comparing the NN Voronoi cells with its GFC equivalent local capturing regions of each individual sample shown in Fig. 4b. The cells in the Voronoi diagram are transformed into long stripes stretching along field vectors. The reason for this is that the data field, being a result of the global influence of all samples, tend to establish major radial directions around data mass centres affected locally on smaller scale.

3 Electrostatic model with repelling force

So far data have only been attracted to each other which was a consequence of the same charge carried by all the samples regardless of the class labels. In such an approach the information about class labels is in a sense wasted during the data matching process and just in the final stage of classification are the class labels from the sources passed statically to the captured samples. 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. There are many ways of incorporating repelling force into the field definitions. One of unsupervised solutions is to adopt intermolecular potential containing both short-range attracting and long-range repelling components in the distance definition (2). However, in this section we consider a maximally supervised approach where labels of the training dataset are exploited during the action of the field. The most suitable physical model seems to be the electrostatic field, where opposite charges attract each other and charges of the same sign repel. To adopt this rule for 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 a positive potential of the same absolute value, triggering a repelling force.

3.1 Labels partition estimation

The major problem with the electrostatic data field metaphor 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 a testing sample is decomposed into smaller subsamples labelled by all classes present in the field generated by the training data. The partition between the sizes of subsamples has to be controlled by any external classifier producing soft outputs. The Parzen density estimation based classifier seems ideal in this case. It helps determine whether the field mechanism applied on top of the classifier improves its performance or not.

Given the training set $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ with labels vector $L_s = \{l_1, \dots, l_n\}$ where $l_i \in \{1, \dots, C\}$ the objective is to evaluate the labels partition matrix $P^{[N \times C]}$ for the testing set $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$.

According to the Parzen-window approach [2] an estimate of the data density in point \mathbf{y}_j can be obtained by calculating:

$$p(\mathbf{y}_j) = p_j = \frac{1}{n} \sum_{i=1}^n \frac{1}{V_n} \varphi\left(\frac{\mathbf{y}_j - \mathbf{x}_i}{h_n}\right) \quad (10)$$

where V_n is a window volume and φ is a specific window function and h_n is a smoothing parameter. In our model we use the Gaussian window function defined by:

$$\varphi(u) = \frac{1}{2\pi} e^{-u^2/2} \quad (11)$$

and apply the leave-one-out maximum likelihood estimation of the optimal smoothing parameter h_n . Let \mathbf{p}_k denote a vector of Parzen density estimates for the testing set Y but generated only by training samples from the k th class, $k \in \{1, \dots, C\}$. To obtain the class partitions in the form of pseudo-probabilities all densities \mathbf{p}_k are transformed using standard classification mapping:

$$\mathbf{p}_k^N = \frac{1}{1 + e^{-\mathbf{p}_k}} \quad (12)$$

and scaled to sum up to unity:

$$\mathbf{p}_k^S = \frac{\mathbf{p}_k^N}{\sum_{i=1}^C \mathbf{p}_i^N} \quad (13)$$

Given the probability vectors \mathbf{p}_k^S the required labels partition matrix is simply:

$$P = [\mathbf{p}_1^S, \dots, \mathbf{p}_C^S] \quad (14)$$

3.2

Electrostatic field definition

Given the labels partition matrix P , note that classification of the unlabeled testing set is readily available by just finding the maxima of each row corresponding to the index of the class with maximum Parzen density in considered position \mathbf{y}_j .

Let L_X denotes a vector of labels from the training set. The definition of the potential (3) in the new version including repelling aspects takes the following form:

$$U_j = \sum_{i=1}^n \left(\underbrace{\frac{\sum_{k \neq L_i} p_{jk}}{r_{ij}}}_{\text{repelling}} - \underbrace{\frac{p_{jL_i}}{r_{ij}}}_{\text{attraction}} \right) \quad (15)$$

making use of the normalisation: $\sum_{k=1}^C p_{jk} = 1$ potential (15) can be rewritten simply by

$$U_j = \sum_{i=1}^n \frac{1 - 2p_{jL_i}}{r_{ij}} \quad (16)$$

which implies immediately the new definition of the field vectors as:

$$\mathbf{E}_j = \sum_{i=1}^n \left[(1 - 2p_{jL_i}) \frac{\mathbf{y}_j - \mathbf{x}_i}{r_{ij}^3} \right] \quad (17)$$

Note that the numerator of the potential definition (16) can be both positive and negative depending on the class partial memberships stored in matrix P . In the presence of many classes, regardless of their topology, the values of the matrix P will naturally decrease to share the evidence with many classes. Effectively the potential would grow positive which means that repelling force will dominate the field landscape. In our model the data still have to slide down the potential towards the source samples. To satisfy this 'potential well' condition 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:

$$\sum_{j=1}^N U_j = 0 \quad (18)$$

To achieve this the goal equation (16) should be parameterised and solved with respect to the regularisation coefficient s as in the following:

$$\sum_{j=1}^N U_j = \sum_{j=1}^N \sum_{i=1}^n \frac{1 - sP_{jL_i}}{r_{ij}} = 0 \quad (19)$$

In the model we use the bisection method to find a numerical estimation of the parameter s . Note that parameter s has a meaningful interpretation, as the value $s - 1$ says in general how many times the attractive interaction should be stronger to compensate for the excess repelling interaction 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 the physical electrostatic field we will refer to the presented method as an *electrostatic field classifier* or EFC for short.

3.3

Classification example

As an example we take the same artificial dataset that was used in Sect. 2.4. The optimisation process of zeroing the total potential (19) resulted in the regularisation coefficient $s = 3.356$. Figure 5 illustrates the shape of the electrostatic field and is compared to the gravity field discussed in Sect. 2. Clearly the improved class separation has been achieved, which resulted in smoother boundaries and reduced the effect of deeply inclining class stripes, commonly observed for GFC classifier boundaries.

4

Diversity

Diversity among classifiers is the notion describing the extent, to which classifiers vary in data representation, concepts, strategy and so on. However, this multimodal perception of diversity results in a simple effect observed on the outputs of classifiers: they tend to make errors for different input data. This phenomenon has been shown to be crucial for effective and robust combining methods [5, 10, 11]. The diversity can be measured in a variety of ways [5, 7, 10], but the most effective turned out to be the measures evaluating direct disagreement on errors among

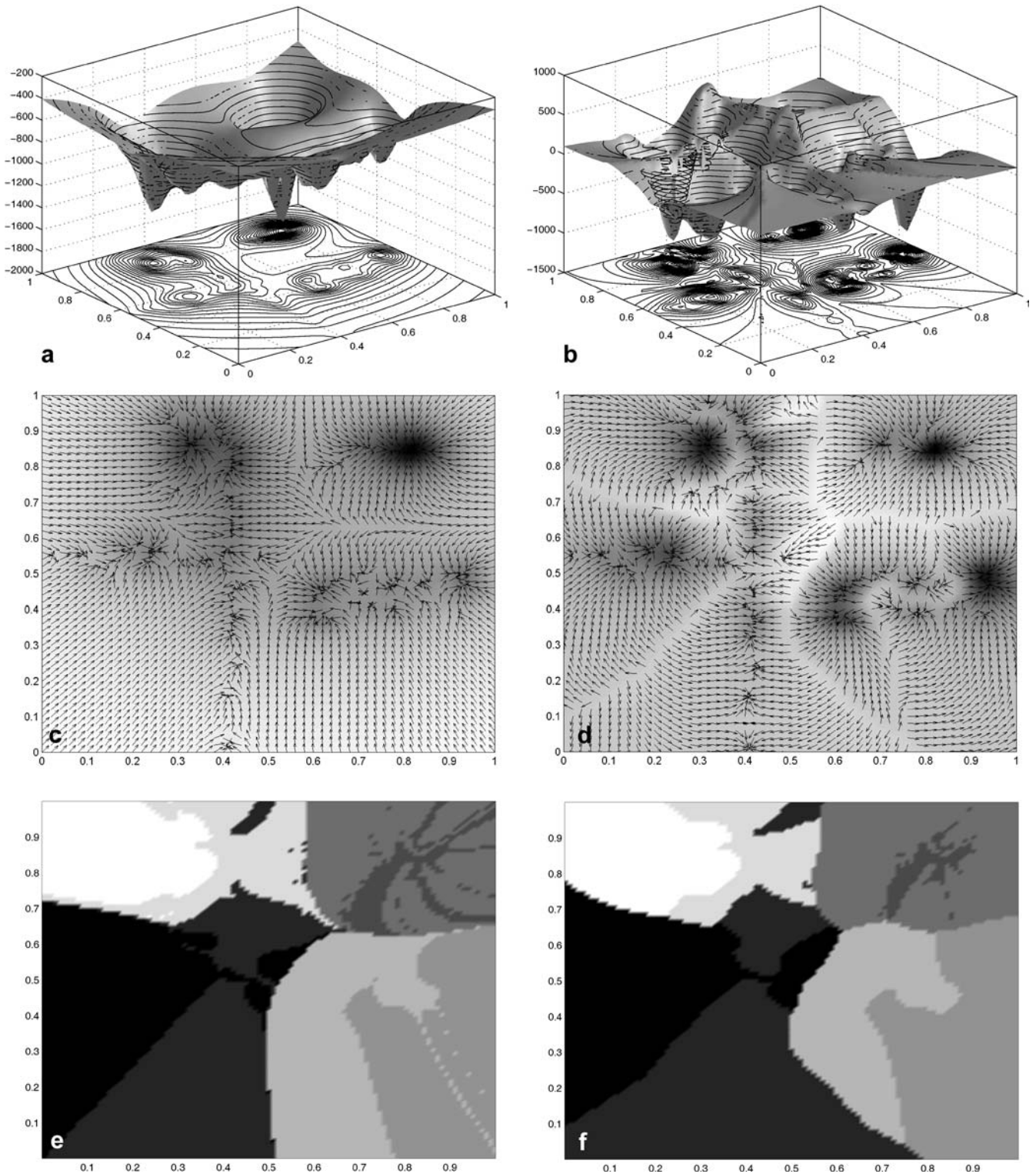


Fig. 5a-f. Visual comparison of the GFC and EFC classifiers for 8-class 2-D artificial dataset. **a, b** 3-D images of the potential for GFC and EFC respectively, **c, d** Vector plots of the field intensities

for corresponding classifiers, **e, f** Decision boundary diagrams for GFC and EFC classifiers respectively

classifiers [5, 7–9]. In [7] we investigated the usefulness and potential applicability of a variety of pairwise and non-pairwise diversity measures operating on binary outputs (correct/incorrect). As we concluded in [7] and further developed in [8] and [9], the key to a successful diversity measure is its asymmetry to the output swap

which reflects the greater focus on measuring error coincidences.

For the purpose of this paper we will be using a double fault (DF) measure capturing 2nd order error coincidences or in other words estimating the probability of both errors for a pair of classifiers. In our analysis [7], the DF measure

turned out to be the best among analyzed pairwise measures in terms of correlation to majority voting. Recalling the definition of the DF measure, the idea is to calculate the ratio of the number of samples misclassified by both classifiers n^{11} to the total number of samples n :

$$F = n^{11}/n \quad (20)$$

Using this simple measure one can roughly assess the average diversity between all pairs of classifiers which potentially indicates the team strength of the classifiers when combined.

5 Experiments

The presented static field based classification methods require a number of evaluation procedures. First of all we check the performance of both models over typical real and artificial datasets and compare them against other classifiers. Secondly, as mentioned above, we based the static field classification on the original concept of the physical fields hoping to achieve a satisfactory level of diversity with other classifiers. In the second part of the experiments both field models are evaluated in terms of pairwise diversity (double-fault measure) with typical classifiers from the machine learning domain. It is important for a good classifier to meet both these conditions on a satisfactory level to be successfully used in combining schemes. In the final part of the experimental section we investigate the dependence of the performance on model parameters and indicate some general observations.

Experiment 1

In this experiment we used 6 well-established artificial and real datasets to check an individual performance of both the GFC and EFC classifiers and compared them to the performances of another 13 typically used classifiers. Table 1 shows the details of the datasets picked and a list of all the classifiers examined 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 200 samples for training and 500 for testing, drawn at random. Classification runs have been repeated 10 times for different random splits. Table 2 shows averaged individual performances of classifiers from this experiment. Although the performances of the GFC and EFC classifiers are never the best, they consistently remain among the top classifiers. As mentioned above, to become a competitive candidate for combining, in addition to good individual performance, a classifier needs to exhibit sufficient diversity properties. The next experiment provides rough estimates of the pairwise diversity among all pairs of classifiers from the considered pool of 15 classifiers.

Experiment 2

The purpose of this experiment was to evaluate the diversity for all pairs of classifiers and then look at average diversity properties of each individual classifier. As mentioned in Sect. 4, double fault seems was shown to be a good, yet simple, indicator of the diversity exhibited among classifiers. From the previous experiment we used

Table 1. Description of datasets and classifiers used in the experiments

Dataset	Code	#Samples	#Feat	#Class	Classifiers
Iris	iri	150	4	3	klc PCA based linear classifier log Logistic linear classifier fis Minimum least square linear classifier ldc Normal densities-based linear classifier nmc Nearest mean linear classifier
Conetorus	cnt	400	2	3	qdc Normal densities-based quadratic classifier qua Quadratic classifier
Gaussians	gau	250	2	2	fsv Pseudo Fisher support vector classifier knn K-nearest neighbour classifier par Parzen density based classifier
Azizah	azi	291	8	20	sub Subspace classifier tre Decision tree classifier
Segment	seg	2310	19	7	lmn Feed-forward neural network classifier rbn Radial basis neural network classifier bpn Back-propagating neural network classifier
Shuttle	shu	58000	9	7	gfc Gravity field classifier efc Electrostatic field classifier

Table 2. Individual performances of classifiers applied to 6 considered datasets described in Table 1

	klc	log	fis	ldc	nmc	qdc	qua	fsv	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	5.89	4.59	11.28	4.28	4.56	4.33
cnt	27.25	25.68	26.54	27.23	29.44	20.32	19.86	17.05	17.73	16.25	60.02	18.78	18.60	31.83	18.98	18.89	16.69
syn	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

classifier outputs hardened to a binary form of 1-correct, 0-error. As a result, for each dataset we obtained a binary matrix with 15 columns – corresponding to the number of classifiers involved. Calculation of the DF measures for all pairs of classifiers resulted in 15 by 15 matrices of diversity values for each dataset. For presentation purposes the results are shown graphically in the form of diversity diagrams in Fig. 6. The coordinates of each small square correspond to the indices of classifiers for which the DF measure is calculated. Note that the diagrams are diagonally symmetrical as there is effectively only $(M^2 - M)/2$ pairs for M classifiers for which unique DF measures are to be calculated. The shade of the squares reflects the magnitude of the DF measure values. The lower the DF measure, the lighter the square and the more diverse the corresponding pair of classifiers. To get an overall single-value indicator of diversity properties of individual classifiers, we averaged the DF measures between a considered

classifier and all remaining classifiers. This is shown numerically in Table 3 and also graphically on the diagonal of the diagrams. Both, diagrams from Fig. 3 and averaged results from Table 3 show very good diversity properties of both GFC and EFC classifiers. Only the Fisher support vector classifier out of 15 classifiers shows on average better diversity, which is mainly due to its outstanding individual performances. It is quite interesting that the EFC classifier, though using Parzen density estimation, is more diverse than the actual Parzen classifier. It confirms that the original concept of the field applied for classification substantially improves the diversity properties of the classifier. To support this finding, the analysis of the last 2 columns in Table 3 reveals that even the GFC classifier built purely on the field concept shows better diversity properties for most of the datasets and only quite bad performance for the difficult Azizah dataset, which substantially decreases its average diversity value.

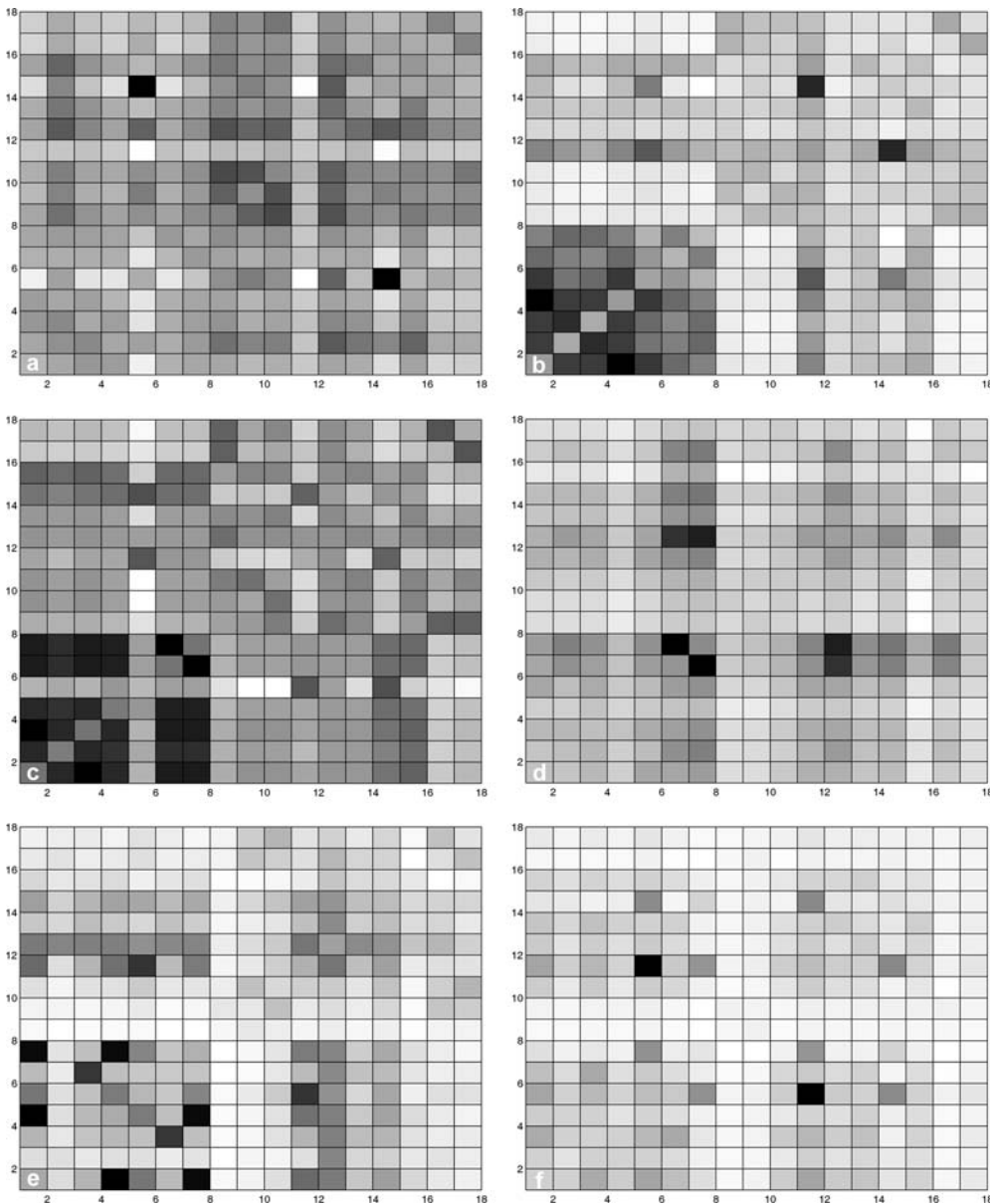


Fig. 6. Diversity diagrams obtained for 6 considered datasets in the order as shown in Table 1. The coordinates of each small square correspond to the indices of classifiers for which DF measure is calculated. The shade of the squares reflects the magnitude of DF measure. The lower is the DF measure the lighter is the square and the more diverse is corresponding pair of classifiers. The column-wise average DF measures are shown on the diagonals of the diagrams

Table 3. Averaged values of DF measures between individual classifiers and all the remaining classifiers. The DF values have been expressed as a percentages of the occurrences of pairwise coincident errors to the total number of samples

	klc	log	fis	ldc	nmc	qdc	qua	fsv	knn	par	sub	tre	lmn	rbn	bpn	gfc	efc
iri	1.78	2.43	2.02	1.80	1.94	1.78	2.07	2.62	2.44	2.56	1.46	2.74	2.24	2.08	2.34	1.92	1.96
cnt	15.32	14.32	14.13	15.31	14.71	13.32	12.71	10.87	10.69	10.68	15.22	10.43	11.65	11.65	12.25	10.39	10.38
syn	11.01	10.61	11.01	10.77	8.50	10.97	10.97	9.38	9.22	9.40	8.91	9.91	9.34	9.73	10.05	8.59	8.92
azi	30.07	31.55	31.54	24.82	32.46	43.79	46.46	23.33	24.38	27.64	32.93	40.67	30.07	33.94	21.23	31.51	23.74
seg	7.17	4.94	6.36	6.14	6.69	5.89	6.93	4.30	5.15	5.66	7.00	7.56	6.18	6.83	4.89	4.86	5.35
shu	6.84	4.85	6.45	4.95	8.88	5.80	4.90	2.45	2.92	4.57	8.95	5.31	5.73	5.52	5.06	2.64	3.48
Ave	12.03	11.45	11.91	10.63	12.19	13.59	14.01	8.82	9.13	10.085	12.41	12.77	10.86	11.62	9.30	9.98	8.97

Table 4. Individual error rates of various configurations of GFC classifier. The results obtained for *Conetorus* dataset using 10 random splits (50% for training, 50% for testing)

Potential definition	Rational $U_j(x) = -\sum_{i=1}^N 1/r_{ij}^a$	Exponential $U_j(x) = -\sum_{i=1}^N e^{-ar_{ij}}$						
Parameter a	0.1	1	2	5	10	20	50	100
Error rate	16.67	15.54	15.76	15.82	-	16.62	16.54	16.06

Experiment 3

In the last experiment we investigated a parametrical variability of the presented classifier. Recalling the force definition (7), the only parameter of the field having a potential influence on the classification results is the type of distance appearing in the potential definition (1). For both field based models we investigated the changes in the field landscape, decision boundaries and the classifier performance observed as a result of different distance functions. We used the 2-dimensional *Conetorus* dataset for visualisation of the imposed changes.

First we applied the GFC classifier with different powers of the distance appearing in the denominator of the potential definition (3). Additionally, we examined a simple exponential distance function with one parameter as an alternative definition of potential. Table 4 shows all configurations of the GFC examined in this experiment, as well as individual performances obtained on the testing set of the *Conetorus* dataset. Visual results including field images and class boundaries are shown in Fig. 7 for the rational potential definition and Fig. 8 for the exponential definition. For both functions the results depict a clear meaning of the parameter a . Namely it accounts for the balance between local and global interactions among the samples. The larger the value of a , the more local the field, so that virtually only the nearest neighbours influence the field in a particular point of the input space. For smaller a the field becomes more global and below a certain critical value some training samples are no longer able to intercept any testing samples. Technically this occurs when a single source cannot curve the geometry strongly enough to create a closed enclave of higher potential around itself. In realistic situations this condition can always be satisfied with a rational potential if the lower limit on distance d is small enough. For the exponential definition of the potential the condition of potential wells is much easier to break as its value is limited for $r = 0$. Effectively, as shown in Table 4 for $a \leq 10$, some testing samples have been trapped in the global potential minimum in the centre which is not occupied by any training sample. For such

cases the GFC classifier is unable to classify all the samples. It is interesting that the classifier performance is not changing much for different values of parameter a . The reason for this phenomenon is that the relative changes of the field are not changing rapidly with varying parameter a , in spite of the fact that the absolute values of potential are hugely affected. Due to the above it is safer to use larger than smaller values of the parameter a , just to ensure the condition of potential wells is met. Moreover, the optimal value of a seems to be a function of the number of samples and its optimization could be included in the classifier design.

Enriched by the above findings for the EFC classifier we inspected the changes using only a relational potential definition for 3 different values of the parameter a as shown in Table 5. The results of the obtained error rates and visual illustrations of the obtained electrostatic fields conform to the previous findings for the GFC model. The only difference is that performance is improved by around 1% and as concluded in Sect. 3 the smoother decision boundaries are obtained as a result of the Parzen density estimation routine for partial labelling of the testing set.

6 Conclusions

In this paper, we introduced a novel non-parametric classification method based on the static data field adopted from the physical field phenomena. The meaning of training data has been reformulated as sources of a central static field with the negative potential increasing with the distance from the source. Two field models have been developed with an explicit analogy to the attracting gravity

Table 5. Individual error rates of various configurations of EFC classifier. The results obtained for *Conetorus* dataset using 10 random splits (50% for training, 50% for testing)

Potential definition	$U_j(x) = \sum_{i=1}^N (1 - 2P_{jL_i})/r_{ij}^a$		
Parameter a	0.1	1	2
Error rate	14.74	14.21	15.06

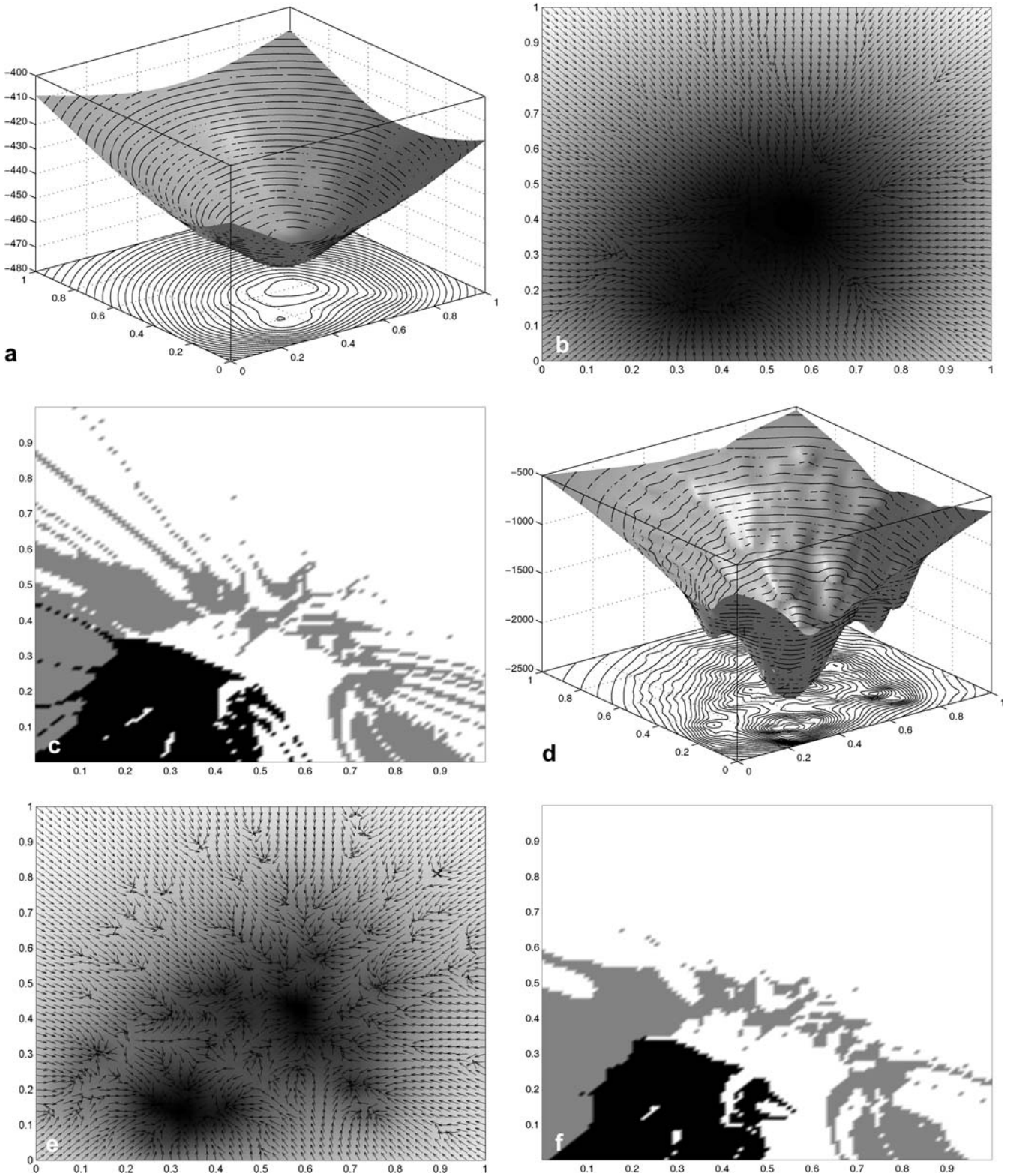


Fig. 7a-l. Visualisation of the GFC classifier for different values of parameter a in the rational potential definition of the form: $U_j(x) = -\sum_{i=1}^N 1/r_{ij}^a$. **a-c** Potential plot, vector field intensity

and class boundaries plots for $a = 0.1$, **d-f** $a = 1$, **g-i** $a = 2$, **j-l** $a = 5$

field and the electrostatic field, with both attracting and repelling forces. In the gravity model, the attracting force among the data emerges as a gradient of a specific complex potential landscape resembling the joint potential wells

built around the training data – *field sources*. The classification process has been proposed as a gradient descent translocation of the unlabelled testing sample ultimately forced to approach one of the sources and inherit its label.

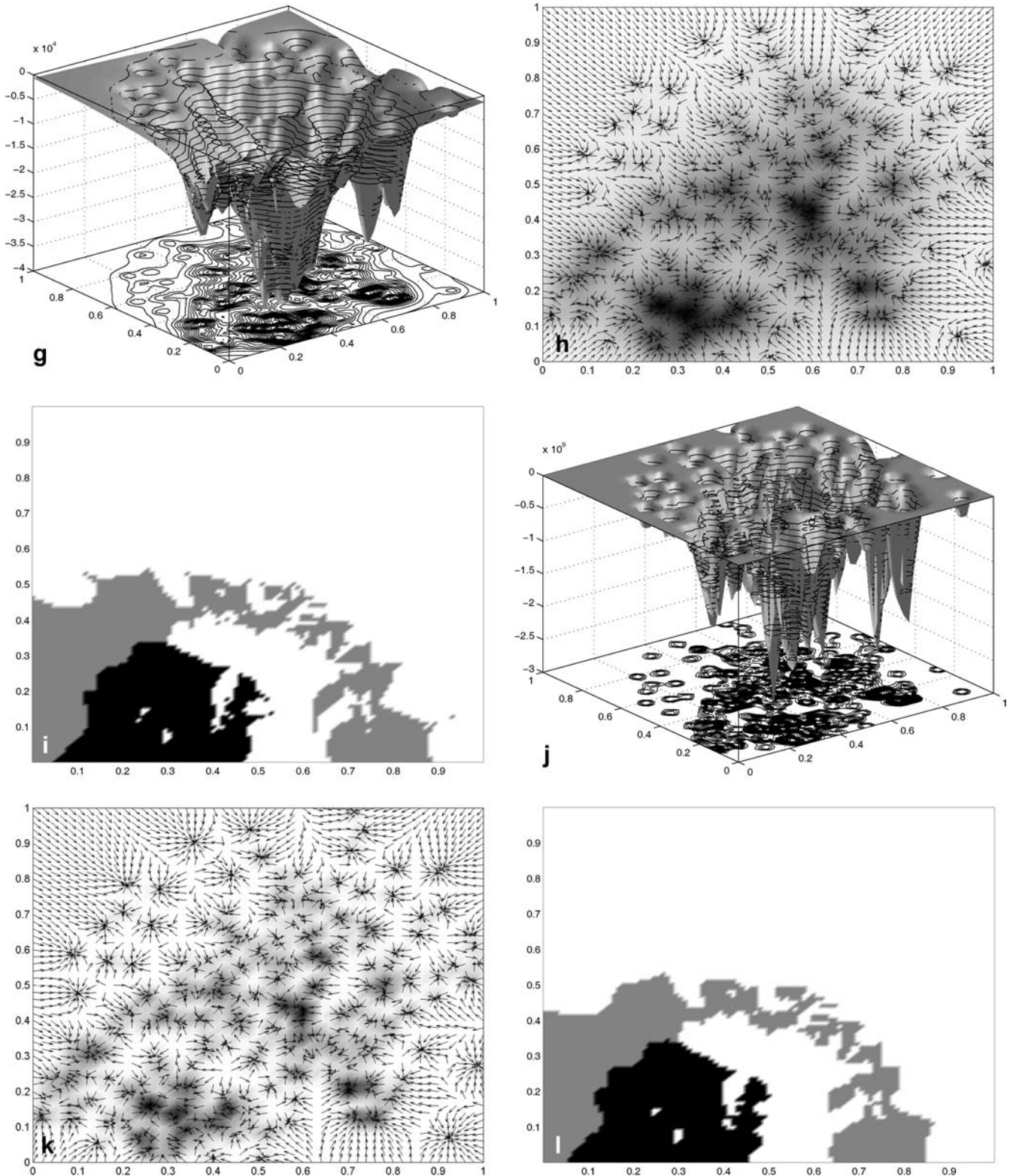


Fig. 7a-l. Contd.

We noted that the gravity field classifier represents a purely data-driven model where the information about the data labels is completely ignored. In the electrostatic field model the attractive forces adopted from the gravity model have been applied only to samples from the same class, whereas repelling force has been added as an interaction affecting samples from different classes. In this model,

samples have been split into sub-samples carrying charges of each class partitioned and normalised according to the Parzen class density estimation. The overall balance between attracting and repelling interaction is controlled by the zero-potential rule. Both gravity field (GFC) and electrostatic field classifiers (EFC) have been implemented using an efficient matrix formulation suitable for

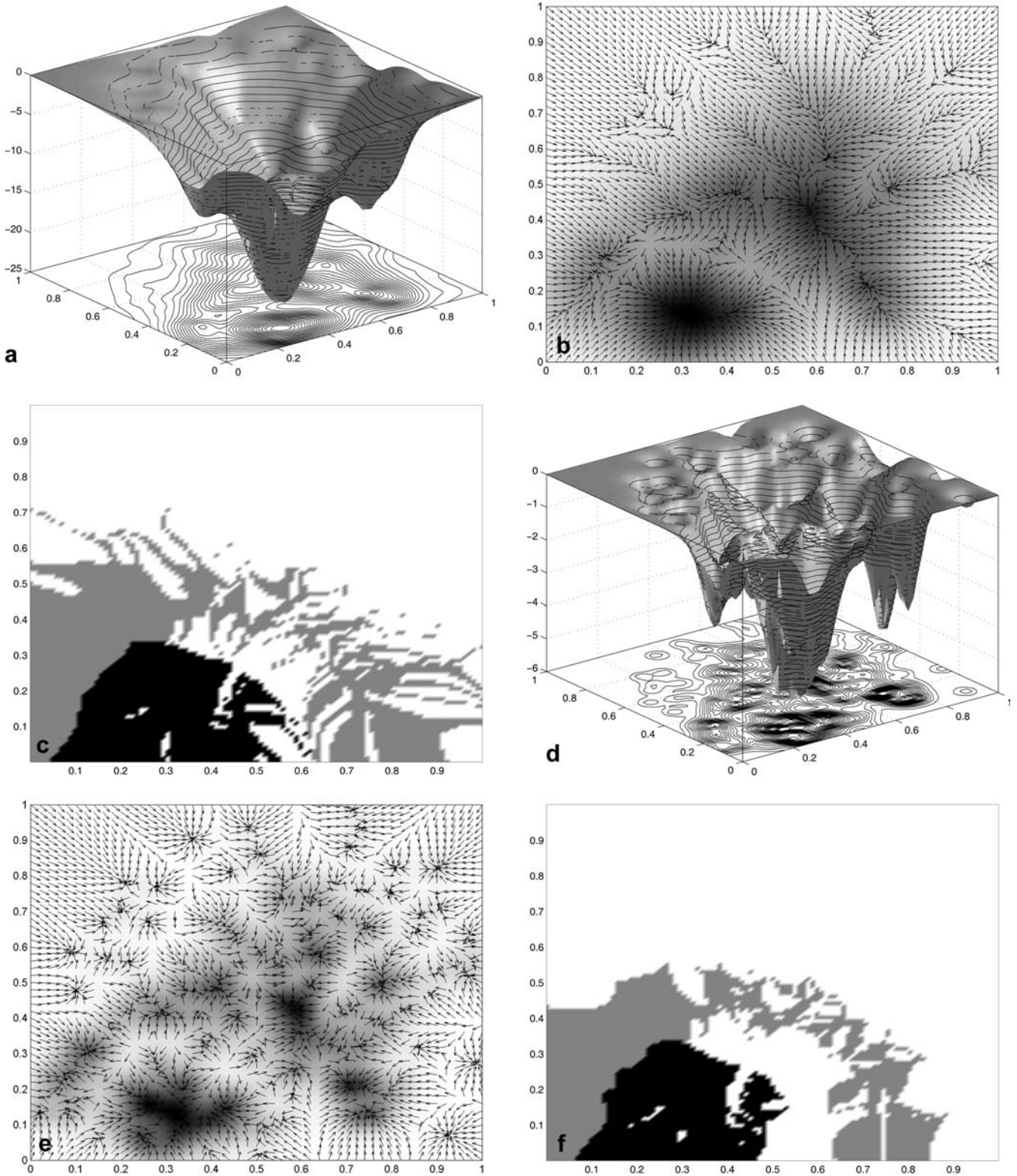


Fig. 8a-f. Visualisation of the GFC classifier for different values of parameter a in the exponential potential definition of the form: $U_j(x) = -\sum_{i=1}^N e^{-ar_{ij}}$, **a-c** Potential plot, vector field intensity and class boundaries plots for $a = 20$, **d-f** $a = 50$

parallel processing and analysis by mathematical software like Matlab. Extensive graphical content has been used to depict different geometrical interpretations of the field classifiers as well as fully visualise the classification process.

The presented classifiers have been evaluated in a number of ways. An individual performance has been examined on 6 established datasets and compared to 13 other high performing classifiers. The results showed relatively average performance of the GFC classifier and

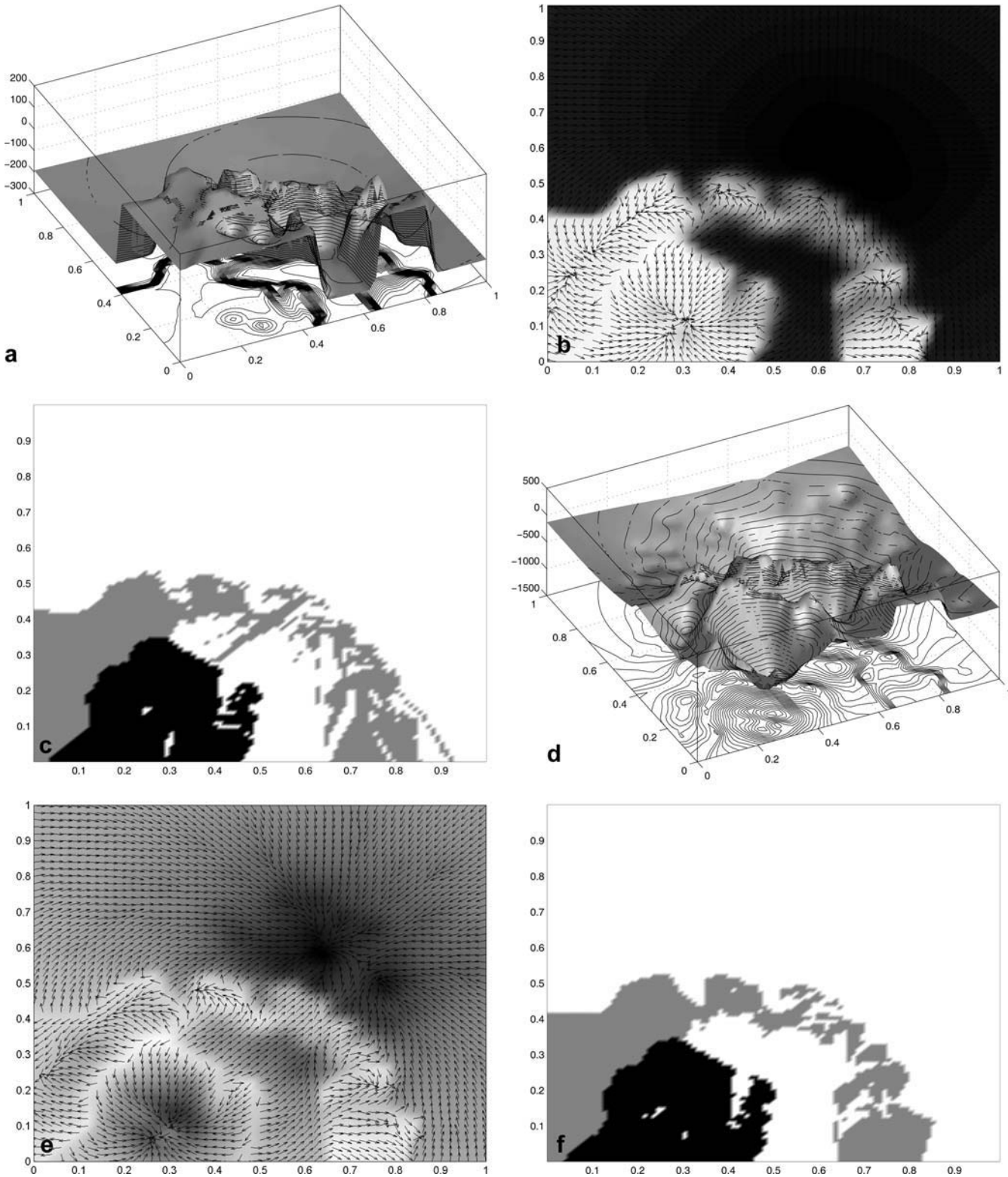


Fig. 9a-i. Visualisation of the EFC classifier for different values of parameter a in the rational potential definition of the form: $U_j(x) = \sum_{i=1}^N (1 - 2P_{jL_i})/r_{ij}^a$. **a-c** Potential plot, vector field intensity and class boundaries plots for $a = 0.1$. **d-f** $a = 1$, **g-i** $a = 2$

quite good performance of the EFC classifier if applied individually. The electrostatic field model has better performance and generalisation properties due to improved class separation and smoother boundaries. The analysis of the diversity properties of the novel classification models

reveal very diverse patterns of classification outputs especially for the GFC. Compared to 13 other established classifiers over 6 datasets EFC was the 2nd and GFC was the 4th most diverse classifier as evaluated by average pairwise double fault measure. These results seem

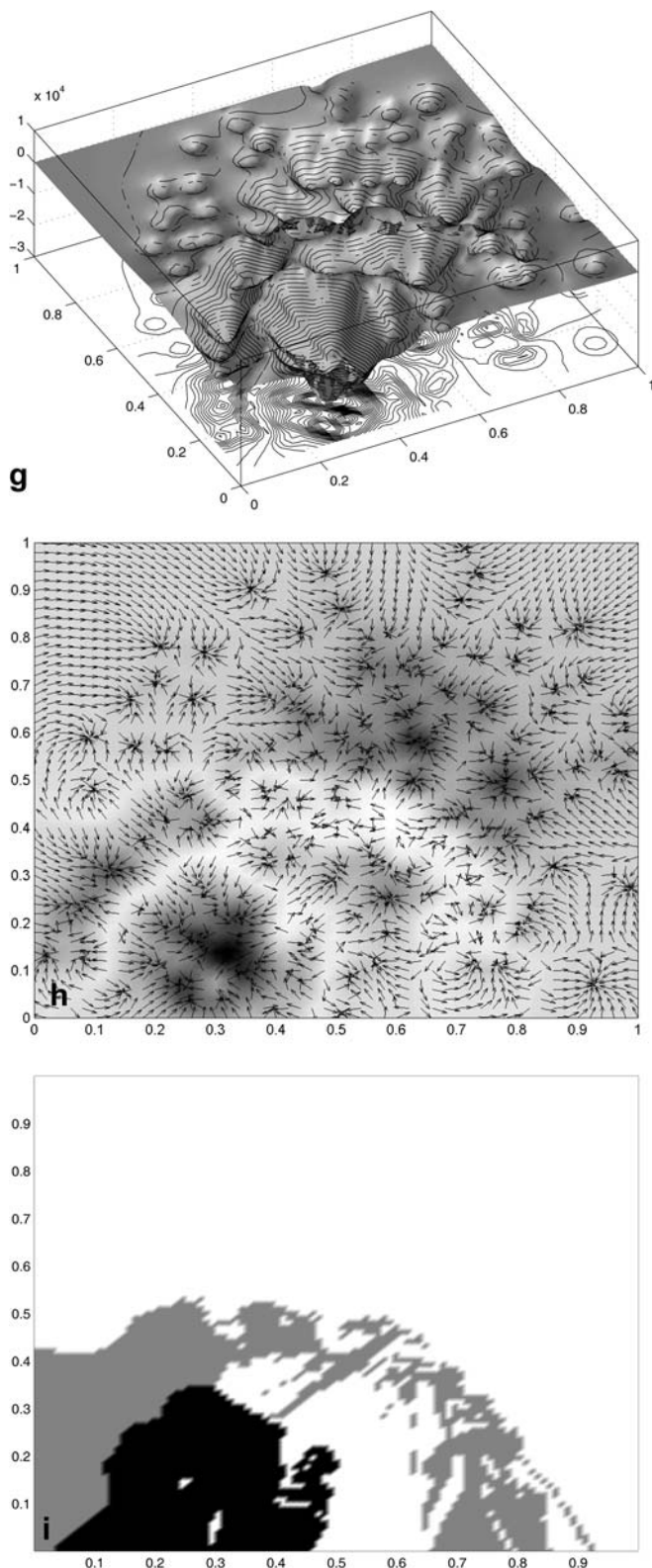


Fig. 9a-i. Contd.

especially encouraging in the case when the corresponding performance ratings are 3 and 8 for the same set of classifiers. The substantial improvement of the diversity related to individual performance is clearly observed.

Additional improvement of both the performance and the diversity properties may be achieved by careful parameter settings, examples of which are shown in the experimental section. However, the differences between outputs of the GFC and EFC classifier suggest that for the highly diverse GFC model to reach consistently better performance its diversity had to be sacrificed. At the same time it seems that diversity is largely correlated with the classifier performance as the best performing classifiers usually remain among the most diverse ones if diversity is measured using DF or other error coincidence based measures. Various types of fields have also been examined within both the GFC and EFC models. The conducted experiments suggested the use of local fields for the best performance as well as for boundary invariability, but further experiments are required for fuller interpretation of these results.

The properties mentioned above as well as the results from the presented experiments allow the proposed models to be considered as an alternative to non-parametric approaches for classification, and are particularly useful in combination with other classifiers.

References

1. **Bezdek JC** (1999) Fuzzy models and algorithms for pattern recognition and image processing. Kluwer Academic, Boston
2. **Duda RO, Hart PE, Stork DG** (2001) Pattern classification. John Wiley & Sons, New York
3. **Hochreiter S, Mozer MC** (2000) An Electric Approach to Independent Component Analysis. Proc. of the Second International Workshop on Independent Component Analysis and Signal Separation, Helsinki, pp. 45–50
4. **Klir GJ, Folger TA** (1988) Fuzzy Sets, Uncertainty, and Information. Prentice-Hall International Edition
5. **Kuncheva LI, Whitaker CJ** (2001) Ten measures of diversity in classifier ensembles: limits for two classifiers. IEE Workshop on Intelligent Sensor Processing, Birmingham 10/1–10/6
6. **Principe J, Fisher III, Xu D** (2000) Information Theoretic Learning. In Haykin S (Ed.) Unsupervised Adaptive Filtering. New York
7. **Ruta D, Gabrys B** (2001) Analysis of the correlation between majority voting errors and the diversity measures in multiple classifier systems. Proc. of the SOCO/ISFI'2001 Conference, ISBN: 3-906454-27-4, paper #1824–025, Paisley, UK
8. **Ruta D, Gabrys B** (2002) New measure of classifier dependency for multiple classifier systems. In: Proc. of the 3rd Int. Workshop on Multiple Classifier Systems, Cagliari, Italy, pp. 127–136
9. **Ruta D, Gabrys B** (2002) Set analysis of coincident errors and its applications for combining classifiers, to appear as chapter in the upcoming book: Pattern Recognition and String Matching, Kluwer Academic
10. **Sharkey AJC, Sharkey NE** (1997) Combining diverse neural nets, The Knowledge Engineering Review 12(3): 231–247
11. **Sharkey AJC** (1999) Combining artificial neural nets: ensemble and modular multi-net systems. Springer-Verlag, Berlin Heidelberg New York
12. **Torkkola K, Campbell W** (2000) Mutual information in learning feature transformations. Proc. of International Conference on Machine Learning, Stanford CA
13. **Torkkola K** (2001) Nonlinear feature transforms using maximum mutual information?, Proc. of IJCNN'2001, Washington DC, USA
14. **Zurek WH** (1989) Complexity, Entropy and the Physics of Information. Proc. of the Workshop on Complexity, Entropy, and the Physics of Information. Santa Fe