Electrostatic Field Classifier for deficient data

Marcin Budka[1] and Bogdan Gabrys[2]

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

[2] bgabrys@bournemouth.ac.uk

Summary. This paper investigates the suitability of recently developed models based on the physical field phenomena for classification problems with incomplete datasets. An original approach to exploiting incomplete training data with missing features and labels, involving extensive use of electrostatic charge analogy, has been proposed. Classification of incomplete patterns has been investigated using a local dimensionality reduction technique, which aims at exploiting all available information rather than trying to estimate the missing values. The performance of all proposed methods has been tested on a number of benchmark datasets for a wide range of missing data scenarios and compared to the performance of some standard techniques. Several modifications of the original electrostatic field classifier aiming at improving speed and robustness in higher dimensional spaces are also discussed.

1 Introduction

Physics of information has recently emerged as a popular theme. The research on quantum computing and the concept of 'it from bit' [16] have motivated researchers to exploit physical models for design of learning machines.

One example of such learning model is the Information Theoretic Learning framework derived in [9]. The framework enables online manipulation of entropy and mutual information by employing the concept of Information Potential and Forces. Using higher order statistics of the probability density functions, the common 'Gaussianity' assumption has been lifted, resulting in efficient methods for problems like Blind Source Separation [6], nonlinear dimensionality reduction [14] or training of MLPs without error backpropagation [9].

Another example are the Coulomb classifiers described in [7] which form a family of models based on analogy to a system of charged conductors, trained by minimizing the Coulomb energy. The classifiers are in fact Support Vector Machines but the physical analogy leads to novel types of them, comparable or even superior to standard SVMs.

The dynamic physical field analogy forms a basis of a universal machine learning framework derived in [12]. The Electrostatic Field Classifier is a model of particular interest. By exploiting a direct analogy with the electrostatic field, the approach treats all data samples as particles able to interact with each other. EFC has proven to be a robust solution [11] featuring relatively high level of diversity, which made it suitable for classifier fusion.

The missing data problem is typical for many research areas. There are many reasons why data might be missing and many ways of dealing with it. For probabilistic models the Expectation Maximization algorithm [3] is the commonly used approach, which allows to obtain maximum likelihood estimates of model parameters (e.g. learn the parameters of mixture models). Maximum likelihood based training neural networks has been described in [15], where Radial Basis Networks are used to obtain a closed-form solution instead of expensive integration over the unknown data. The missing data problem has been mainly treated in the statistics literature. The ideas and various types of missingness introduced in [10] are still in use today and the multiple imputation method is considered as state of the art alongside the EM algorithm [13].

A different approach not requiring imputation of missing values based on hyperbox fuzzy sets has been presented in [5]. The architecture of the General Fuzzy Min-Max Neural Networks naturally supports incomplete datasets, exploiting all available information in order to reduce a number of viable alternatives before making the classification decision. The GFMM networks are also able to quantify the uncertainty caused by missing data.

All physically inspired models mentioned above have been designed to handle complete data patterns only. This paper describes an extension of EFC to support incomplete data as well as some other improvements of it.

This document is structured as follows. Section 2 contains a description of data field models which are the subject of this paper. Section 3 has been devoted to the missing data problem, giving a brief overview of the issue, describing some traditional approaches to missingness and a data field model approach used in experiments. The experimental results have been given in section 4, while the conclusions can be found in section 5.

2 Data field model classifiers

2.1 Gravity Field Classifier (GFC)

The simplest, attractive field model has been based on the gravity field [12]. The model treats all data patterns as charged particles and assumes a static field with all sources (training patterns) fixed to their initial positions.

Denoting by X a set of n training samples and by Y a set of N test samples the potential generated by the field source x_i in the point y_j is given by:

$$V_{ij} = -cs_i \frac{1}{r_{ij}} \tag{1}$$

where c is the field constant, s_i is the source charge and r_{ij} is some distance measure between x_i and y_j. For simplicity and for the sake of conformity with the physical model, Euclidean distance has been chosen.

The superposition of individual contributions of all training samples defines the field in any particular point of the input space:

$$V_j = -c \sum_{i=1}^{n} \frac{s_i}{r_{ij}} \tag{2}$$

and the overall potential energy in point y_j is then given by:

$$U_j = s_j V_j = -cs_j \sum_{i=1}^{n} \frac{s_i}{r_{ij}} \tag{3}$$

Assuming that all samples have the same, unit charge, s_i and s_j can be dropped from the equations and the force exerted on y_j by the field (negative gradient of U_j) can be calculated as:

$$F_j = -c \sum_{i=1}^{n} \frac{y_j - x_i}{r_{ij}^3} \tag{4}$$

The training dataset uniquely identifies the field, which implies that no training is required and all calculations are performed during classification. The potential results in a force able to move an unlabelled test sample to finally meet one of the fixed field sources and share its label. Only force directions are followed, taking a small, fixed step d at a time (which makes the field constant c irrelevant). This allows to avoid problems in the vicinity of field sources, as $r_{ij} \to 0$ implies $F_j \to \infty$. The field is then recalculated and the procedure repeats until all testing samples approach one of the sources at a distance equal to or lower than d and are labelled accordingly. Due to the fact that d is fixed in all dimensions, the data should be rescaled to fit within the $0-1$ range. The lower bound of the distance is also set to d, to avoid division by zero and overshooting the sources. Note, that all possible trajectories end up in one of the field sources, the space is thus divided into distinct regions representing classes.

2.2 Electrostatic Field Classifier (EFC)

During classification the GFC does not take advantage of training data labels until the very end of the procedure. This information is thus wasted. The way to exploit it is to use the electrostatic field analogy, by introducing a repelling force into the model, so that samples from the same class would attract each other, while samples from different classes – repel. Since class label of a test sample is not known, it cannot directly interact with the field. To facilitate this interaction, each testing sample must be decomposed into a number of subsamples belonging to one of the target classes. This can be achieved by using some density estimator (e.g. Parzen window). In order not to introduce additional parameters, those partial memberships can be assigned in proportion to the GFC potential of all classes in the test point.

Denoting by L the vector of labels of the field sources and by V_j^k the potential generated by k^{th} of C classes in point y_j, the partial membership p_{jk} of the sample y_j in the k^{th} class is given by:

$$p_{jk} = \frac{|V_j^k|}{\sum_{i=1}^{C} |V_j^i|} \tag{5}$$

while the overall potential in point y_j can be calculated as:

$$V_j = \sum_{i=1}^{n} \left(\frac{\sum_{k \neq L_i} p_{jk} - p_{jL_i}}{r_{ij}} \right) = \sum_{i=1}^{n} \frac{1 - 2p_{jL_i}}{r_{ij}} \tag{6}$$

The resultant force calculation formula then becomes:

$$F_j = \sum_{i=1}^{n} \left[(1 - 2p_{jL_i}) \frac{y_j - x_i}{r_{ij}^3} \right] \tag{7}$$

Note however, if there are more than two classes, repelling force may dominate the field, as it would come from multiple classes, while the attracting force would come from only one. According to [12], to restore the balance between repelling and attracting forces it is sufficient to satisfy the following condition:

$$\sum_{j=1}^{N} V_j = \sum_{j=1}^{N} \sum_{i=1}^{n} \frac{1 - qp_{jL_i}}{r_{ij}} = 0 \tag{8}$$

by estimating a value of the regularization coefficient q. This coefficient controls the balance between the total amount of attracting and repelling force in the field but as discussed in the following section, the condition given above may in fact not be sufficient. This can result in some test samples being repelled by the field, which would prevent the algorithm from converging.

The classification process follows the same rules as in the case of gravity field model. As expected, classification performance and generalisation properties are better than in the case of the simpler model due to improved class separation and smoother decision boundaries [12].

2.3 Improvements of the Electrostatic Field Classifier

In some applications the EFC tends to suffer from a number of issues:

- Excess of repelling force. The regularization coefficient q satisfying Eq. 8 is often too small to restore the force balance and the algorithm diverges. The formula also makes the field landscape dependant on the test data, thus a new way to estimate q was needed. An artificial test set is generated by placing samples in the corners of the field (Fig. 1(a)) and the value of q is estimated so that forces in all test locations point into the field. If the training dataset is representative, this ensures that no test sample will escape during classification. The coefficient can then be stored and reused for other test sets.
- Distance concentration. When the number of dimensions grows roughly above 5 or 6, EFC tends to diverge or produce a very high classification error. This behavior can be explained by the properties of Euclidean distance, which is a natural choice for 2 or 3 dimensional spaces but looses its discriminative power as the number of dimensions grows [1]. Under a broad set of conditions the mean value of the Euclidean distance distribution grows with dimensionality, while the variance remains approximately constant, which results in the ratio of distances to the nearest and farthest neighbour tending to 1 [4]. This 'concentration' issue is usually slower for norms of lower order and thus by relaxing the L_p-norm assumption that $p \geq 1$, a family of concentration resistant similarity measures called 'fractional distances' has been derived in [1] and incorporated into EFC to ensure robustness in high dimensional spaces.

Some other improvements regarding adaptive simulation step size manipulation and label assigning procedure have also been included in the current implementation of EFC.

3 Handling incomplete data

3.1 The missing data problem

The missing data problem is typical for many research areas. The reason of missingness has important implications, thus data can be divided into [10]:

- Missing Completely At Random (MCAR), if the probability that the particular feature is missing is not related to the missing value or to the values of any other features. The best kind of missingness one can hope for [8].
- Missing At Random (MAR), if the probability that the particular feature is missing is not related to its own value but is related to the values of other features. There is no way to test if MAR holds but usually false assumption of MAR may have only minor influence on the result [13].
- Missing Not At Random (MNAR), if the probability that some feature is missing is a function of this feature value. The missingness should be somehow modelled in this case but the model for missingness is rarely known, which makes the whole procedure a difficult and application specific task [8].

EFC is a purely data-driven approach, thus the type of missingness does not directly influence its operation, although it can influence the results. This dependency is however not investigated here and MCAR data was assumed.

3.2 Basic approaches to missingness

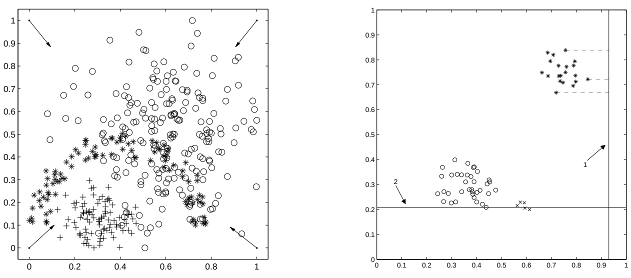
There exist some basic approaches to the missing data problem, based on editing. In statistical inference, usefulness of most of them is limited to a number of specific cases [8, 13] but their performance within the data field framework is in some cases quite reasonable. The techniques are:

- Casewise deletion, which simply excludes incomplete data samples from further analysis. If their number is relatively small, this technique performs surprisingly well. This method has been extensively used in the experiments in section 4 as a base for performance comparison with other methods.
- Pairwise deletion, which also ignores missing data but instead of dropping incomplete samples the approach uses only the features which are present. A similar method forms a basis of the data field specific approach to classification of incomplete patterns, as described in the following subsection.
- Mean substitution, which replaces all missing features with appropriate mean values. Although commonly criticized in the literature [8, 13], as shown in section 4, in many cases class conditional mean imputation turns out to perform very well in conjunction with the data field model specific approach.

3.3 EFC approach to missingness

Various modifications were required to facilitate handling of deficient data within the EFC framework. The modifications include distance calculation, force calculation and label assignment routines and are discussed below.

- Classification of incomplete data. To exploit all available information, EFC acts on the incomplete test sample working only in available dimensions – the feature space dimensionality is locally reduced. As a result, distance and force calculations take place in the reduced feature space and the test sample is only able to move within it (Fig. 1(b)). The pattern can also no longer simply share the class of the nearest source as it might lead to ambiguity. Instead, a soft, probability-like output is produced, proportional to the class conditional density for the current position of the test sample in the reduced feature space.



(a) Regularization coefficient

(b) Incomplete test data

Fig. 1. Representation and classification of incomplete test patterns

- Learning from incomplete data. The missing feature scenario can be addressed by reintroducing the charge concept – if the charge is allowed to vary between field sources, the incomplete training patterns can be exploited by an intelligent charge redistribution mechanism. The algorithm starts with assigning a unit charge to all training samples. It then examines each incomplete sample in turn and redistributes its charge among all complete patterns from the same class in proportion to their distance. Thus the closer the complete sample is, the more charge it receives. After all incomplete patterns are processed they are dropped and the remaining samples become field sources. The missing labels scenario has been addressed in 2 different ways: by treating unlabelled samples as gravity field sources (GFC fallback) or by redistributing the charge among all complete field sources regardless of their class.

4 Experimental results

The experiments include evaluation of classification error for various levels of missing data on a number of benchmark datasets from the UCI Machine Learning Repository [2]. All recognition rates given have been averaged over 10 runs with randomly removed features and labels (MCAR). The results for the following scenarios are provided: (1) deficient test data, (2) deficient training data, (3) deficient both types of data with missing labels and (4) deficient both types of data with labels given.

Table 1. Iris dataset results for scenarios (1), (2) and (3)

deficiency type/level[3]	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
(1) local dim reduction	95.5	94.7	94.4	93.8	91.6	89.4	88.4	84.0	83.8	82.4	77.5
(1) mean imputation	95.5	92.8	90.2	85.2	80.5	76.9	71.3	67.5	61.1	55.8	50.4
(2) charge redistribution	95.4	95.7	94.8	94.7	95.0	94.6	92.6	89.1	87.0	85.0	85.2
(2) mean imputation	95.4	95.3	95.1	94.7	94.5	95.4	94.0	93.5	93.0	92.0	92.3
(2) casewise deletion	95.4	95.3	93.9	91.8	90.7	90.6	86.2	78.1	74.8	72.5	75.6
(3) best methods[4]	95.1	94.2	93.7	92.3	90.9	86.9	85.9	82.4	80.9	71.1	73.5
(3) casewise deletion	95.1	93.5	93.0	89.7	85.8	80.8	72.8	69.5	73.2	69.3	71.2

[3] 0 for complete and 1 for maximally incomplete data (one feature left for each object)

[4] combination of best performing methods from previous experiments

Table 2. Wine dataset results for scenarios (1), (2) and (3)

deficiency type/level	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
(1) local dim reduction	96.9	95.2	94.8	93.8	91.9	90.3	88.1	84.4	79.2	72.9	71.4
(1) mean imputation	96.9	93.1	89.6	84.5	78.9	75.9	70.5	66.0	57.5	53.6	51.8
(2) charge redistribution	96.4	94.8	92.7	86.1	83.3	82.1	82.8	80.6	82.1	83.2	81.5
(2) mean imputation	96.4	96.3	96.7	96.9	96.1	96.1	96.1	96.1	96.4	94.4	92.7
(2) casewise deletion	96.4	87.9	72.3	62.9	72.0	72.9	73.3	70.8	74.9	73.6	76.4
(3) best methods	96.5	94.7	94.1	91.8	91.5	87.8	83.6	78.6	67.9	60.8	58.0
(3) casewise deletion	96.5	86.5	69.0	66.6	70.1	68.8	68.1	65.7	59.9	57.7	56.8

The results have been given in Tables 1 – 4. As it can be seen for all datasets, local dimensionality reduction significantly outperforms mean imputation in the incomplete test data scenario (1). The advantage margin steadily grows up to more than 27 percentage points (Table 1), as the deficiency level increases.

The situation changes in favor of mean imputation, when the training dataset has missing features (2). The performance of the method is always the best and in some cases (Table 1 and 2) does not drop almost at all even at the highest deficiency level. This phenomenon

Table 3. Ionosphere dataset results for scenarios (1), (2) and (3)

deficiency type/level	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
(1) local dim reduction	85.2	86.5	86.3	86.5	86.1	86.7	87.6	86.7	85.9	84.0	80.7
(1) mean imputation	85.2	84.5	82.0	78.3	75.7	72.6	69.3	68.1	66.6	65.4	64.9
(2) charge redistribution	85.6	71.4	60.9	58.7	61.8	62.6	62.7	63.1	61.6	62.6	60.7
(2) mean imputation	85.6	86.6	86.6	86.2	86.5	86.8	86.8	85.7	84.4	82.2	71.5
(2) casewise deletion	85.6	71.6	58.5	55.4	57.9	57.1	59.6	57.5	59.2	57.2	57.4
(3) best methods	85.8	85.5	86.3	85.4	84.5	84.3	81.2	77.1	72.3	62.7	58.6
(3) casewise deletion	85.8	72.3	56.9	57.1	55.9	55.8	56.8	57.1	56.9	55.7	56.3

Table 4. Segment dataset results for scenarios (1), (2) and (3)

deficiency type/level	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
(1) local dim reduction	92.0	91.5	90.9	90.3	88.4	87.8	84.8	79.9	72.0	59.6	43.4
(1) mean imputation	92.0	86.4	80.2	72.5	64.2	56.3	48.5	40.8	33.4	25.3	18.5
(2) charge redistribution	92.3	87.5	70.6	62.7	65.7	64.2	63.5	62.4	63.3	63.0	59.6
(2) mean imputation	92.3	91.8	90.9	89.9	88.8	88.3	87.5	87.0	85.3	83.6	75.4
(2) casewise deletion	92.3	81.5	52.3	52.4	57.0	55.5	58.3	54.5	56.3	57.6	55.6
(3) best methods	91.9	90.5	87.8	85.4	82.2	78.2	72.8	65.8	56.9	44.5	34.8
(3) casewise deletion	91.9	78.4	51.3	54.2	53.1	55.2	51.4	47.3	44.7	39.6	31.5

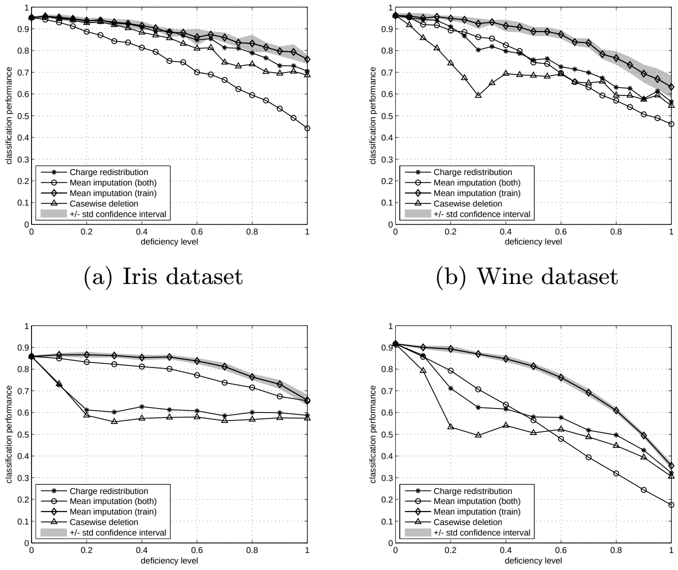
must be however credited to good class separation of the Iris and Wine datasets, rather than to the EFC model.

In the most difficult scenario (3) the proposed model always outperforms simple casewise deletion, although at high deficiency level the performance margin is rather modest. Note however, that the performance drop of the former is much smoother (for casewise deletion the lowest recognition rate can be reached at 0.2 deficiency already) and even at the maximum deficiency level allowed by the model, it is still better than random guessing.

The results for scenario (4) have been depicted in Fig. 2. Notice the superiority of class conditional mean imputation and local dimensionality reduction combination and its smooth performance decay. For Iris and Wine datasets, the latter method is almost entirely responsible for the performance drop, as discussed before.

5 Conclusions

The underlying physical model of EFC appears well suited for incorporation of various missing data handling routines. The approaches investigated in this paper, although criticized in the statistical literature, perform quite reasonably in conjunction with a non-parametric, physical field model. The performance of those methods appears not as problem dependant as one would expect – mean imputation, which intuitively should perform well only for certain types of datasets with well separated classes, is the best method for dealing with deficient training data in most experiments. There also emerges a pattern of what approach is most likely to produce the best results with a particular type of data missing: (1) local dimensionality reduction for incomplete test data, (2) class conditional mean imputation for training data with missing features and (3) charge redistribution for missing labels. A combination of above methods provides good recognition rates even for the most difficult scenarios, for both low and moderate deficiency levels. A peculiarity of the model is its limited sensitivity to removal of unlabelled training data – casewise deletion often performs only slightly worse than charge redistribution.



(a) Iris dataset

(b) Wine dataset

(c) Ionosphere dataset

(d) Segment dataset

Fig. 2. Classification performance for deficient test and training data

References

1. C. Aggarwal, A. Hinneburg, and D. Keim. On the surprising behavior of distance metrics in high dimensional space. *Lecture Notes in Computer Science*, 2001.
2. A. Asuncion and D. Newman. UCI machine learning repository, 2007.
3. A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via EM Algorithm. *Journal of the Royal Statistical Society*, 39(1), 1977.
4. D. Francois, V. Wertz, and M. Verleysen. Non-Euclidean metrics for similarity search in noisy datasets. *Proceedings of the European Symposium on Artificial Neural Networks*, pages 339–334, 2005.
5. B. Gabrys. Neuro-fuzzy approach to processing inputs with missing values in pattern recognition problems. *Journal of Approximate Reasoning*, 30(3), 2002.
6. K. Hild, D. II, and J. Príncipe. Blind Source Separation Using Renyis Mutual Information. *IEEE SIGNAL PROCESSING LETTERS*, 8(6), 2001.
7. S. Hochreiter, M. Mozer, and K. Obermayer. Coulomb classifiers: Generalizing support vector machines via an analogy to electrostatic systems. *Advances in Neural Information Processing Systems*, 15:545–552, 2003.
8. W. Outhwaite and S. Stephen P Turner. *Handbook of Social Science Methodology.* SAGE Publications Ltd., 2007.
9. J. Principe, D. Xu, and J. Fisher. Information theoretic learning. *Unsupervised Adaptive Filtering*, pages 265–319, 2000.
10. D. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.
11. D. Ruta and B. Gabrys. Physical field models for pattern classification. *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, 8(2):126–141, 2003.
12. D. Ruta and B. Gabrys. A Framework for Machine Learning based on Dynamic Physical Fields. *Natural Computing Journal on Nature-inspired Learning*, 2007.
13. J. Schafer and J. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002.
14. K. Torkkola. Feature extraction by non parametric mutual information maximization. *The Journal of Machine Learning Research*, 3:1415–1438, 2003.
15. V. Tresp, S. Ahmad, and R. Neuneier. Training neural networks with deficient data. *Advances in Neural Information Processing Systems*, 6:128–135, 1994.
16. W. Zurek. *Complexity, Entropy and Physics of Information.* Westview, 1989.