Independent Component Analysis in Knowledge Discovery in Databases Process: A Fuzzy and Genetic Approach

Luis Miguel Marín Trechera[1], Francisco Álvarez González[2], and
Jorge Ollero Hinojosa[3]

[1] Dpto. Estadística e I.O., Escuela Superior de Ingeniería, Universidad de Cádiz,
11003 Cádiz, Spain
luis.marin@uca.es

[2] Dpto. Estadística e I.O., Fac. de Ciencias de la Educación, Universidad de Cádiz,
11510 Puerto Real (Cádiz), Spain
francisco.alvarez@uca.es

[3] Dpto. Estadística e I.O., Facultad de Ciencias, Universidad de Cádiz,
11510 Puerto Real (Cádiz), Spain
jorge.ollero@uca.es

Abstract. Feature extraction plays a fundamental role in the KDD and Data Mining process. There are many algorithms for mining data based in Principal Component Analysis (PCA), a powerful statistical tool which is identical to the Karhunen-Loeve transform for pattern recognition. Independent Component Analysis (ICA) is a recently developed technique based on the assumption of statistical independence between the components that acts as a remedy to the limitations of PCA.

In this paper we describe some applications of ICA in the KDD process and in the Data Mining step of this process. We propose a fuzzy method to quantify the information of a linear combination of input data and a genetic algorithm to find the components with the optimal values of such measure.

1 Introduction

Independent Component Analysis is a powerful statistical and computational tool for revealing underlying factors from multivariate statistical data. It can play a fundamental role in KDD process.

The structure of this paper is as follows: In the next section we describe the classical KDD paradigm. Section 3 presents an overview on Independent Component Analysis. In section 4 we describe the role that ICA plays in KDD process. Section 5 proposes a fuzzy and genetic approach to finding independent components.

Table 1. Number of pages found by google search engine containing the different terms. Data Mining and KDD are the most used of them.

Term	Number
Data Mining	596000
KDD	172000
Knowledge Discovery	104000
Information Discovery	20300
Data Pattern	11200
Knowledge Extraction	6460
Data Archaeology	1250
Information Harvesting	783

2 Data Mining and KDD

Many terms are used to refer to the art and technology of finding the knowledge hidden from large volumes of raw data. Data mining, information discovery, information harvesting, knowledge discovery in databases, data archaeology, knowledge extraction, or data pattern processing are terms used by statisticians, data analysts, and researches in the AI and machine-learning fields. The definitions are changing and the use of these names depends on the application field and they are sometimes used synonymously. The most used of these terms are Data Mining and KDD, as we can see in Table 1.

The definition on which the KDD community is converging usually places data mining as a particular step in the larger KDD process. KDD is the nontrivial process of identifying valid, novel, potentially used, and ultimately understandable patterns in data. The result of this process is newly acquired knowledge formerly hidden in the data. This new knowledge may then be used to assist in future decision making. The KDD process in interactive and iterative, involving numerous steps. The basic flow of steps can be summarized as follows [1–4]:

1. **Data Selection:** The extraction from a larger data store of only the data that is relevant. This data extraction helps to streamline and speed up the process.
2. **Data Preprocessing:** Data cleaning and preparation tasks that are necessary to ensure correct results. Eliminating missing values in the data, ensuring that coded values have a uniform meaning and ensuring that no spurious data values exist are typical actions that occur during this phase.
3. **Data Transformation:** Finding useful features to represent the data depending on the goal of the task, eliminating unwanted or highly correlated fields so the results are valid.
4. **Data Mining:** The goal of the data mining phase is to analyze the data by an appropriate set of algorithms in order to discover meaningful patterns and rules and produce predictive models. The user can significantly aid the data-mining method by correctly performing the preceding steps.

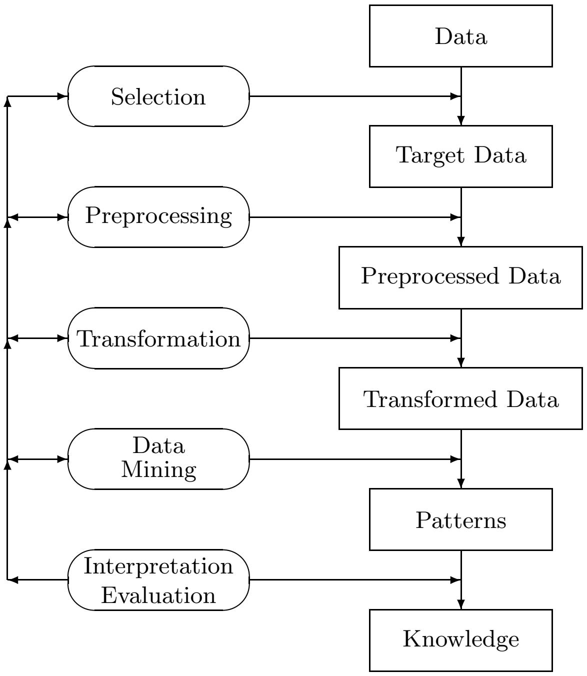

Fig. 1. An overview of the steps in the Traditional KDD Paradigm.

5. **Interpretation and Evaluation:** While data mining algorithms have the potential to produce an unlimited number of patterns hidden in the data, many of these may not be meaningful or useful. This final phase is aimed at selecting those models that are valid and useful for making future decisions.

The KDD process can involve significant iteration and can contain loops between any two steps. This process is illustrated in Fig. 1. The data-mining component has received the most attention in the literature. However, the other steps are as important (and probably more so) for the successful application of KDD in practice. Blind application of single data-mining step can find patterns that appear to be statistically significant but, in fact, are not.

The Data Mining step of the KDD process has two types of goals: *verification* of the user's hypothesis and *discovery* of models (for prediction) and patterns (for description).

3 An overview on Independent Component Analysis

Independent Component Analysis (ICA) is a recently developed technique based on the assumption of statistical independence between the components. ICA can be seen as an extension to principal component analysis. ICA is a much more powerful technique, however, capable of finding the underlying factors or sources when these classic methods fail completely. The goal of linear ICA is to find a linear transform of the observed variables (given by a matrix W) so that the obtained variables (the components) are statistically as independent from each other as possible.

The most used methods in feature extraction are principal component analysis and factor analysis [5–7]. These methods are based on the classical assumption of Gaussianity and only use the information contained in the covariance matrix. There are alternative methods based on higher-order statistics, like projection pursuit [8, 9] , redundancy reduction [10–12], and blind deconvolution [13–16].

The starting point for ICA is the very simple assumption that the components are statistically independent and this fact implies that the independent components must have nongaussian distributions.

There are several methods available for finding independent components, based on different measures of independence or related quantities. In next subsections we briefly review the existing techniques [17–21]. The equivalence of these methods is shown in [22, 23].

3.1 Maximizing Nongaussianity

We must have a quantitative measure of nongaussianity of a random variable such as kurtosis or negentropy. The kurtosis of a random variable, say y, denoted by $\mathrm{kurt}(y)$, is defined by

$$\mathrm{kurt}(y) = E[y^4] - 3(E[y^3])^2 \tag{1}$$

The kurtosis is zero for a gaussian random variable. There are nongaussian random variables that have zero kurtosis, but they can be considered to be very rare. The main problem to measure nongaussianity by kurtosis is that it is very sensitive to outliers. The second important measure of nongaussianity is negentropy. A fundamental result of information theory is that a gaussian variable has the largest entropy among all random variables of equal variance. Based on the entropy of a variable H, the negentropy, denoted by J is defined by

$$J(y) = H(y_{gauss}) - H(y) \tag{2}$$

where y_{gauss} is a gaussian random variable of the same covariance matrix as y. In practice, to maximize negentropy or the absolute valor of kurtosis, a gradient algorithm is used [24].

3.2 Minimization of Mutual Information

Minimization of Mutual Information of the components: Mutual information is the natural information-theoretic measure of the independence of random variables. Using the concept of entropy (or differential entropy for continuous variables), we define the mutual information I between m (scalar) random variables, y_i as follows

$$I(y_1, y_2, \ldots, y_m) = \sum_{i=1}^{m} H(y_i) - H(y) \tag{3}$$

This measure is always non-negative, and zero if and only if the variables are statistically independent. Although minimization of mutual information is equivalent to maximizing the sum of nongaussianities, there are some differences between these two criteria. When we use nongaussianity we force the estimates of the independent components to be uncorrelated.

The independent components are those which have minimum entropy.

3.3 Maximum Likelihood Estimation

One interpretation of maximum likelihood estimation is that we take those parameters values that give the highest probability for the observations. We can apply this principle to finding the coefficients of the mixture matrix with maximum likelihood.

4 ICA in KDD process

ICA can be used in different steps of the KDD process. In preprocessing and transformation steps ICA obtains useful featuring to represent the data. In Data Mining step, ICA can be used for prediction and description goals.

- **ICA in preprocessing and transformation:**
  Later tasks will require a good data representation. Several principles and methods have been developed to find a suitable linear transformation, basic goal of these steps. These methods include principal component analysis, factor analysis, projection pursuit, independent component analysis, and many more. Usually, these methods define a principle that tells which transform is optimal. The optimality may be defined in the sense of optimal dimension reduction, statistical 'interestingness' of the resulting components, simplicity of the transformation, or other criteria, including application-oriented ones. Other approaches are more used than ICA, but this tool must be included in the researchers' arsenal. Many papers which PCA is used can be revisited using ICA approach.
  ICA can be used to extract independent components from different kinds of data, for example, color and stereo images, video data, audio data and hyperspectral data.
- **ICA in Data Mining:**
  In Data Mining step, ICA can be used for prediction and description goals.
  - **Prediction:** Hyvarinen and Bingham shows [25] that when only a subset of the input variables is observed, ICA can be used for regression, i.e. to predict the missing observations. This regression is closely related to regression by a multi-layer perceptron.
  - **Description:** In the space of transformations perfomed by ICA, data can form cluster where discrimination between the different ones can be possible.

5 Proposed Method

All the described methods need to estimate the population measures using a sample (the observed data). The estimation of negentropy, for instance, is very difficult, and higher-order moments have to be used.

We propose a new ICA method based on a fuzzy information measure, and we find the mixture matrix using simulated annealing and genetic algorithms. Let $\mathcal{A}$ be a collection of fuzzy sets $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$ defined over a set S with membership functions μ_i. If $X = \{x_1, x_2, \ldots, x_n\} \subset A$ we, can consider the fuzzy sets over the finite support X, and we can calculate the fuzzy cardinality of each set A_i. There are different cardinality measures in the literature (see [26]). We can define an information measure by:

$$-\sum_{i=1}^{m} \frac{\operatorname{card}(A_i)}{M} \log(\frac{\operatorname{card}(A_i)}{M}) \tag{4}$$

where

$$M = \sum_{i=1}^{m} \operatorname{card}(A_i) \tag{5}$$

This formula is a fuzzy extension of entropy and if all A_i are crisp sets it coincide with classical version given by Shannon. This concept is not the same as fuzzy entropy. Fuzzy entropy is a measure of the degree of fuzziness of a fuzzy set and this extended entropy is a measure of the information of observed data with respect to A_i sets.

If we use the Ralescu's cardinality measure [27], we are really doing a discretization of data, and the obtained measure is equal to entropy of discretized data. If we use as cardinality measure the power of a fuzzy set introduced by De Luca and Termini [28], the formula (4) becomes

$$-\sum_{i=1}^{m} \frac{1}{M} \sum_{j=1}^{n} \mu_i(x_j) \log\left(\frac{1}{M} \sum_{j=1}^{n} \mu_i(x_j)\right) \tag{6}$$

The goal is finding components which minimum value of this measure, The procedure is the same as in section 3.2, but using the alternative information measure. Given a linear combination of the observed data, we use the fuzzy sets shown in figure 2. This election ensures that the proposed measure is invariant for invertible linear transformations.

A genetic algorithm [29, 30] is used to finding mixing matrix elements. GAs operate iteratively on a population of matrices, each of which represents a candidate solution to the problem. The initial population is generated randomly and fitness is given by the information measure. Mutation operator changes the value of a specific element of the matrix and the crossover operator interchanges rows between two matrices.

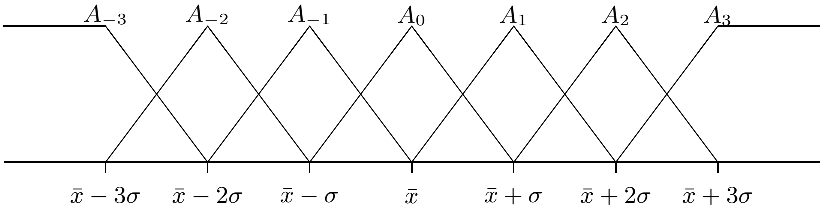

Fig. 2. Membership functions of fuzzy sets.

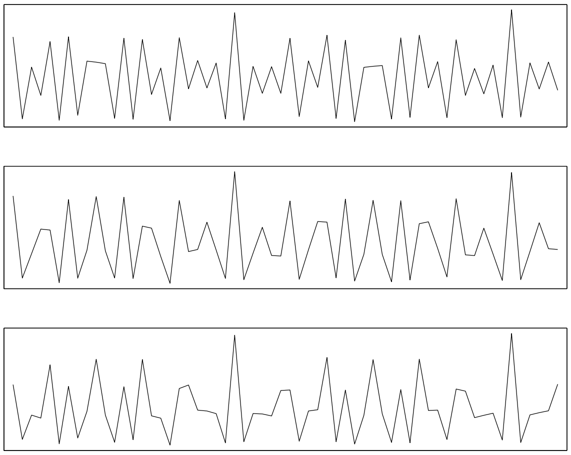

Fig. 3. The observed data

5.1 An Example

The observed data are shown in Fig. 3. Data are strongly correlated. In spite of this, we can apply the method to find the estimates of the original source signals.

Fig. 4 shows the estimates of the original source signals, estimated using only the observed mixture signals. The original signals are not shown, but they are very similar to what the algorithm found. (They could be multiplied by some scalar constants). In this example we use GA to search over the space of all possible mixture matrices with integer elements between -100 and 100.

This method can also be used to find the components with maximum entropy. These components compress the information of the original signals. They are the “principal components” in an information-theoretic sense. In Fig. 5 are shown three different examples of such components. Each of these components compress the information of the original signals: to code one of these components are necessary the same number of bits used to code the three original signals.

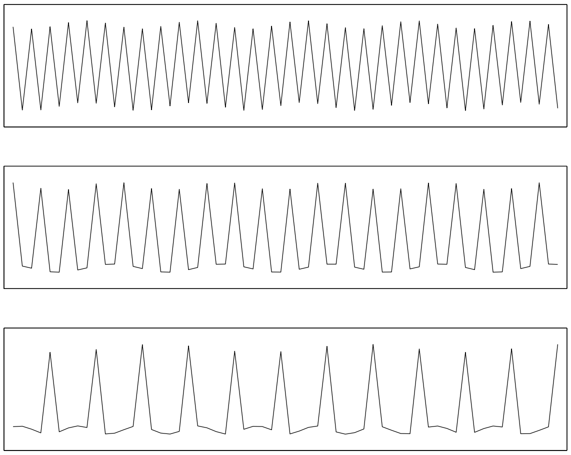

Fig. 4. The independent components found by the method. The original signals were found very accurately

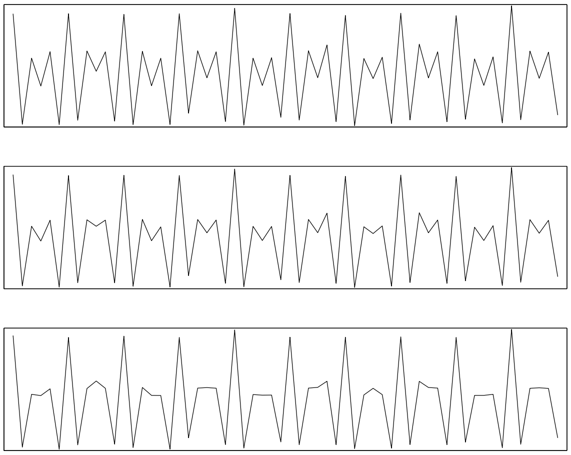

Fig. 5. The maximum entropy components found by the method. The original signals can be obtained from one of these components with a small error

5.2 Conclusions

Independent Component Analysis is a powerful tool that can be used in different steps of the Knowledge Discovery in Database process. In this paper we propose a fuzzy and genetic approach and show a satisfactory application of this method.

In future papers, we will use genetic programming to find nonlinear mixtures of observed data using symbolic regression [31–33].

References

1. Fayyad, Usama, Piatetsky-Shapiro, G. and Smyth, P.: From data mining to knowledge discovery in databases. AI Magazine, 17 (3), 1996, pp. 37–54. `http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf`
2. Fayyad, Usama, Piatetsky-Shapiro, G. and Smyth, P.: Knowledge Discovery and Data Mining: Towards a Unifying Framework. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), p. 82, AAAI Press, 1996. `ftp://ftp.research.microsoft.com/pub/dtg/fayyad/kdd96/fayyad-intro.ps`
3. Fayyad, Usama, and Stolorz, P.: Data mining and KDD Promise and challenges. Future Generation Computer Systems, Vol. 13, Number 2-3, pp. 99-115, November 1997. `http://ftp.research.microsoft.com/pub/dtg/fayyad/FGCS97/UMF-PES.PS`
4. Fayyad, Usama, Piatetsky-Shapiro, G. and Smyth, P.: The KDD Process for Extracting Useful Knowledge From Volumes of Data. Communications of the ACM, 39(11), pp. 27-34, November 1996. `wwwhome.cs.utwente.nl/~mpoel/colleges/dwdm/ACM_artikelen/fayyad2.pdf`
5. M. Kendall. Multivariate Analysis. Charles Griffin&Co., 1975
6. H. H. Harman. Modern Factor Analysis. University of Chicago Press, 2nd edition, 1967
7. I.T. Jolliffe. Principal Component Analysis. Springer-Verlag, 1986.
8. J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. IEEE Trans. of Computers, c-23(9):881-890, 1974.
9. J.H. Friedman. Exploratory projection pursuit. J. of the American Statistical Association, 82(397):249-266, 1987. 6
10. D.J. Field. What is the goal of sensory coding? Neural Computation, 6:559-601, 1994.
11. H. B. Barlow. Single units and sensation: A neuron doctrine for perceptual psychology? Perception, 1:371-394, 1972.
12. H.B. Barlow. What is the computational goal of the neocortex ? In C. Koch and J.L. Davis, editors, Large-scale neuronal theories of the brain. MIT Press, Cambridge, MA, 1994.
13. Haykin, editor. Blind Deconvolution. Prentice-Hall, 1994.
14. S. Haykin. Adaptive Filter Theory. Prentice-Hall International, 3rd edition, 1996.
15. O. Shalvi and E. Weinstein. New criteria for blind deconvolution of nonminimum phase systems (channels). IEEE Trans. on Information Theory, 36(2):312-321, 1990.
16. O. Shalvi and E. Weinstein. Super-exponential methods for blind deconvolution. IEEE Trans. on Information Theory, 39(2):504:519, 1993.

17. A. Hyvärinen. Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. IEEE Transactions on Neural Networks 10(3):626-634, 1999. http://www.cis.hut.fi/aapo/ps/TNN99.pdf
18. A. Hyvärinen and E. Oja. Independent Component Analysis: Algorithms and Applications. Neural Networks, 13(4-5):411-430, 2000. http://www.cis.hut.fi/aapo/ps/NN00.pdf
19. A. Hyvärinen. New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit. In Advances in Neural Information Processing Systems 10 (NIPS*97), pp. 273-279, MIT Press, 1998. http://www.cis.hut.fi/aapo/ps/NIPS97.pdf
20. A. Hyvärinen. Gaussian Moments for Noisy Independent Component Analysis. IEEE Signal Processing Letters, 6(6):145–147, 1999. http://www.cis.hut.fi/aapo/ps/SPL99.ps
21. A. Hyvärinen. Survey on Independent Component Analysis. Neural Computing Surveys 2:94–128, 1999. http://www.cis.hut.fi/aapo/ps/NCS99.pdf
22. J.-F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. IEE Proceedings-F, 140(6):362-370, 1993. http://www.ee.duke.edu/~lcarin/Cardoso_IEE.pdf
23. J.J. Atick. Entropy minimization: A design principle for sensory perception? International Journal of Neural Systems, 3:81-90, 1992. Supp. 1992.
24. A. Hyvärinen, J. Karhunen and E. Oja. Independent Component Analysis. Wiley-Interscience. 2001.
25. A. Hyvärinen and E. Bingham. Connection between multi-layer perceptrons and regression using independent component analysis. Neurocomputing, in press. http://www.cis.hut.fi/aapo/ps/NC02_icamlp.ps
26. M. Delgado, D. Sanchez, M. J. Martin-Bautista and M. A. Vila, A probabilistic definition of a nonconvex fuzzy cardinality, Fuzzy Sets and Systems 126 (2) (2002) pp. 177-190
27. D. Ralescu. Cardinality, quantifiers and the aggregation of fuzzy criteria. Fuzzy Sets and Systems 69 (1995) 355-366
28. A. de Luca, S. Termini: A Definition of a Nonprobabilistic Entropy in the Setting of Fuzzy Sets Theory. Information and Control 20(4): 301-312 (1972)
29. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996)
30. Goldberg, D.E. Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley. (1989)
31. J. R. Koza. Genetic Programming, MIT Press, 1992.
32. Jeroen Eggermont and Jano I. van Hemert. Adaptive Genetic Programming Applied to New and Existing Simple Regression Problems Genetic Programming, Proceedings of EuroGP'2001, LNCS, Vol. 2038, pp. 23-35, Springer-Verlag, 18-20 April 2001. http://www.liacs.nl/~jvhemert/publications/
33. F.Rojas, I.Rojas, R.M.Clemente, C.G. Puntonet. Nonlinear Blind Source Separation Using Genetic Algorithms 3rd International Conference on Independent Component Analysis and Blind Signal Separation. (2001)