Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Mathematical and physical techniques of modeling and simulation of pattern recognition in the stock market

To cite this article: O D Montoya *et al* 2021 *J. Phys.: Conf. Ser.* **2073** 012009

Mathematical and physical techniques of modeling and simulation of pattern recognition in the stock market

O D Montoya[1,2], **D D Narváez**[3], **and C A Ramírez Vanegas**[3]
[1] Universidad Distrital Francisco José de Caldas, Bogotá, Colombia
[2] Universidad Tecnológica de Bolívar, Cartagena de Indias, Colombia
[3] Universidad Tecnológica de Pereira, Pereira, Colombia

E-mail: odmontoyag@udistrital.edu.co

Abstract. The following article presents the analysis through mathematical and physical techniques of large databases, which are very common today, due to the large number of variables (especially in the information and physics industry) and the amount of information that results from a process, therefore an analysis is necessary that allows the Decision in a responsible manner, looking for scientific criteria that support said decisions, in our case a database of the forex system will be taken. Initially, a study and calculation of different measurements between the samples and their characteristics will be carried out to make a good prediction of the data and their behavior using different classification methods inspired by basic sciences. Below is an explanation of the techniques based on the analysis of data components and the correlations that exist between the variables, which is a technique widely used in physical processes to determine the correlations between variables.

1. Introduction

When you have many data, such as those obtained in large databases, it is important to classify them; This is done through pattern recognition. We are then facing an important problem which is the data classification problem, a problem that we will face in this document, which consists of assigning the input pattern to one or more categories, typical examples of classification are voice recognition, ordering of digital images, ordering digital, mostly in the industrial sector and high permeability of the physical sciences, astronomy, materials, etc.

This manuscript shows a detailed analysis of Forex, one of the largest databases that exists in the currency market. To make a judicious analysis, which includes simulation and diagnostic modelling we will initially introduce the concept of a distance matrix to determine how dispersed the samples are among themselves and in this way have an indication of the classification of the data. Next, we will determine the correlation index between the characteristics to determine whether or not there are linear relationships between them, this type of technique is very useful and used in physical systems which are most built models of nature.

To move forward with the analysis, it is important to apply a technique that will allow us to determine a set of data in terms of new uncorrelated variables (components); This technique is known as principal component analysis (PCA). Finally, to fulfill the objective we set ourselves; that is, pattern recognition, we will introduce the confusion matrix, a tool that allows the visualization of the performance of an algorithm used in supervised learning. This technique is used and replicated in most physics and mathematics models, which guarantees optimization and efficiency in data analysis.

2. Methodology

The currency exchange (FOREX) was chosen, which is one of the largest databases of this type. Below are some of the characteristics of the study database. Figure 1 and Table 1 show information from the database to be analyzed, for which there are a total of 3355 samples and 14 characteristics [1].

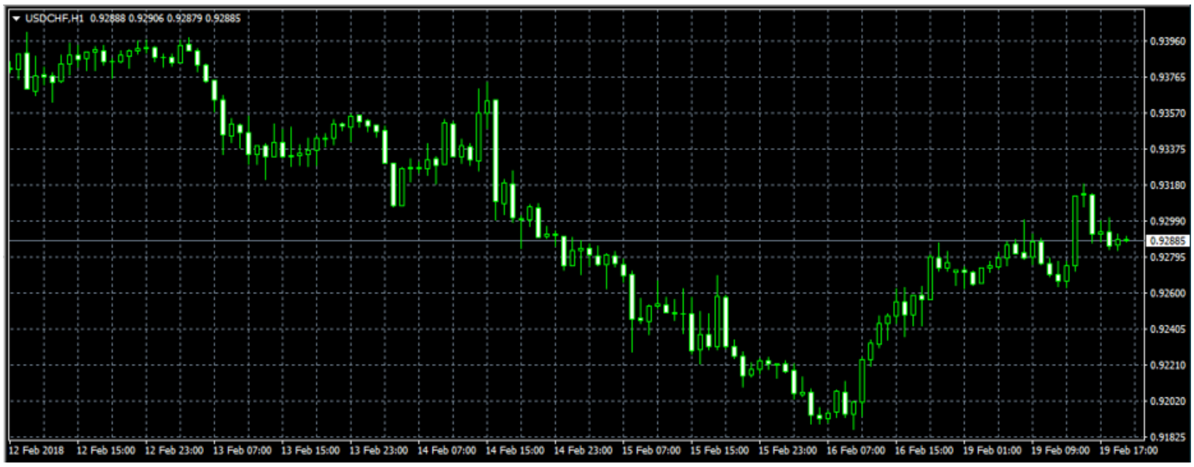


Figure 1. Behavior of the data.

Table 1. Structure of the data.

Date	Rank40	Rank2	StocM	StocS	MacDM	MacDS	Ema14	Ema3	Open-Price	Close-Price
1/08/2017	0.97	0.00	0.00	68.91	63.79	−3.64	0.77	−2.42	−2.03	0.97
2/08/2017	0.97	0.00	0.00	70.01	67.88	−0.85	0.65	−3.39	−1.88	0.97
3/08/2017	0.97	0.00	0.00	57.41	67.16	−3.26	0.10	−5.68	−1.65	0.97
4/08/2017	0.97	0.00	0.00	49.07	63.81	−1.25	−0.77	−6.44	−1.51	0.97
5/08/2017	0.97	0.00	0.00	29.90	55.06	−3.37	−3.04	−6.49	−2.24	0.97
6/08/2017	0.97	0.00	0.00	49.51	51.18	−3.10	−2.00	−3.24	−2.73	0.97
7/08/2017	0.97	0.00	0.00	40.16	45.21	0.13	−1.46	−0.81	−2.30	0.97
8/08/2017	0.97	0.00	0.00	55.58	44.84	2.47	1.08	1.81	−2.80	0.97
9/08/2017	0.97	0.00	0.00	66.27	48.28	10.22	5.90	2.58	−1.51	0.97
10/08/2017	0.96	0.00	0.00	58.08	53.92	5.59	3.21	0.66	−1.65	0.97
11/08/2017	0.97	0.00	0.00	29.89	50.00	−1.52	−1.19	−1.89	−2.78	0.97

The Figure 1 refers to the behavior of the currency exchange (FOREX), which, as can be seen, is quite oscillatory and not very predictable. the white bars that appear there symbolize when the price goes from opening at a low price to closing at a high price and empty bars indicate the opposite. Following is a small sample of the data and the characteristics used in the analysis [2]. Here, we can see that:

- ⇨ Timecurrent: time.
- ⇨ Rank40: range of the last 40 candles.
- ⇨ Rank2: rank of the last 2 candles.
- ⇨ StocM: securities trading of concepts means.
- ⇨ StocS: securities trading of concepts square.
- ⇨ MacDM: moving average convergence divergence mean.
- ⇨ MacDS: moving average convergence divergence square.
- ⇨ Ema3: exponential moving average.
- ⇨ Ema14: exponential moving average.
- ⇨ Ema40: exponential moving average.

- ⇨ Open-Price: Opening price.
- ⇨ Close-Price: Closing price.
- ⇨ OpenClose: Delta of Prices.
- ⇨ Volume.

3. Results

The distance matrix, as the name implies, shows how scattered the samples are to each other and thus have an indication of the classification of the data. Mathematically it can be expressed in the following way [3] Equation (1), we must note that the distance matrix is a concept that comes more from physics than from mathematics, since in this case distance is not used in the strict sense.

$$\mathrm{D} = \begin{bmatrix} \mathrm{d}_{11} & \cdots & \mathrm{d}_{1\mathrm{m}} \\ \vdots & \ddots & \vdots \\ \mathrm{d}_{\mathrm{n}1} & \cdots & \mathrm{d}_{\mathrm{nm}} \end{bmatrix}. \qquad (1)$$

The distance between samples $\mathrm{d}_{\mathrm{ii}} = 0$, so the matrix is Equation (2):

$$\mathrm{D} = \begin{bmatrix} 0 & \cdots & \mathrm{d}_{1\mathrm{m}} \\ \vdots & \ddots & \vdots \\ \mathrm{d}_{\mathrm{n}1} & \cdots & 0 \end{bmatrix}. \qquad (2)$$

Figure 2 represents the distance matrix.

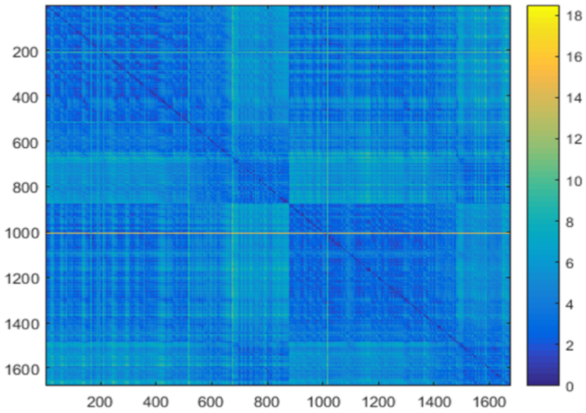


Figure 2. Distance matrix.

Then we will explain from a physical perspective the concepts such as correlations and their relationship between the variables. The correlation is essentially a normalized measure of association or linear covariance between two variables (characteristics) [4]. This measure or correlation index can vary between -1 and $+1$, both extremes indicating perfect, negative, and positive correlations respectively. A correlation value of 0 indicates that there is no linear relationship between the two variables. A positive correlation indicates that both variables vary in the same direction. A negative correlation means that both variables vary in opposite directions. Figure 3 shows the correlation matrix that exists between the characteristics. Geometrically, it can be interpreted by taking two variables (Characteristics) $\mathrm{X}(\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{n})$, Equation (3), and $\mathrm{Y}(\mathrm{y}_1, \dots, \mathrm{y}_\mathrm{n})$, Equation (4), which can be constructed as cantered vectors such as [5,6]:

$$\mathrm{X}(\mathrm{x}_1 - \tilde{\mathrm{x}}, \dots, \mathrm{x}_\mathrm{n} - \tilde{\mathrm{x}}), \qquad (3)$$

$$\mathrm{Y}(\mathrm{y}_1 - \tilde{\mathrm{y}}, \dots, \mathrm{y}_\mathrm{n} - \tilde{\mathrm{y}}). \qquad (4)$$

The correlation coefficient is given by the cosine of the angle formed by the previous vectors Equation (5).

$$r = \cos(\gamma) = \frac{\sum_{i=1}^{N}(x_i-\tilde{x})\cdot(y_i-\tilde{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\tilde{x})^2}\cdot\sqrt{\sum_{i=1}^{N}(y_i-\tilde{y})^2}}. \quad (5)$$

In Figure 4, the relevance of each of the characteristics is shown, likewise, these are ordered from highest to lowest, in order of relevance, and in this case, it has been obtained that the 3 most relevant characteristics are 1; 10 and 11. With the 3 characteristics found, the main components analysis (PCA) is then performed, and a distribution is obtained for the samples shown in Figure 4 [7,8].

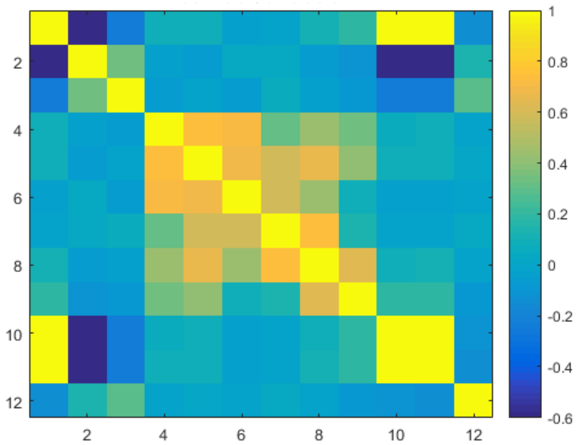


Figure 3. Correlation matrix.

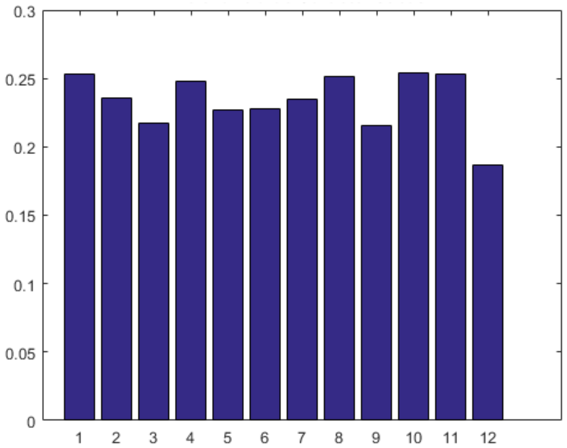


Figure 4. Relevance of characteristics.

Principal component analysis is a technique used to describe a set of data in terms of new uncorrelated variables ("components"). The components are ordered by the amount of original variance they describe, so the technique is useful to reduce the dimensionality of a data set. For its implementation it is necessary to normalize the data to zero mean, calculate the covariance matrix Σ and find the eigenvectors of Σ. For any covariance matrix there is a matrix V called the eigenvector matrix and a diagonal matrix Λ called the eigenvalue matrix [9] Equation (6). It is important to note that the covariance matrix, provides characteristics of the systems through their eigenvalues and eigenvectors, this information comes originally from mathematics and used widely in the physical sciences due to the characterization of the models.

$$\mathrm{PCA}(x) = (X - \mu_x)P. \quad (6)$$

A linear transformation of the normalized data is performed multiplying them by a matrix. That means that if you choose P with great care, you can rotate, scale, or project the data in a vector subspace Equation (7) and Equation (8).

$$\Sigma_x = (x - \mu_x)^T(X - \mu_x), \quad (7)$$

$$\Sigma_x V = V\Lambda, \quad (8)$$

from this analysis, the three-dimensional data is projected by PCA and is shown in Figure 5. The curves presented in Figure 6 represent the success percentages obtained for each of the classification models. The models studied are Linear, Quadratic, Knn, Naive Bayes and SVM [10].

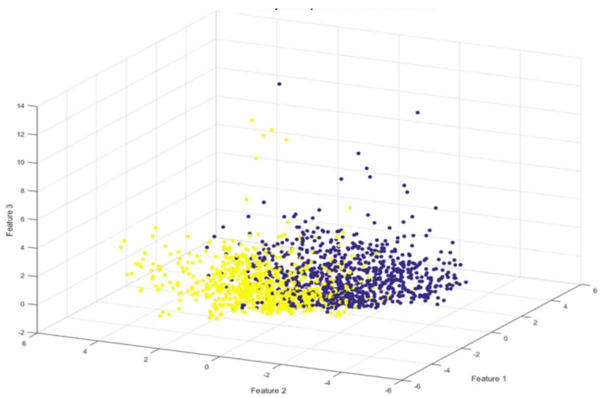


Figure 5. Projected space PCA 3D.

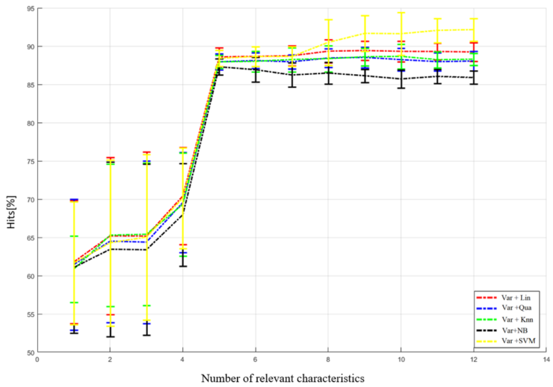
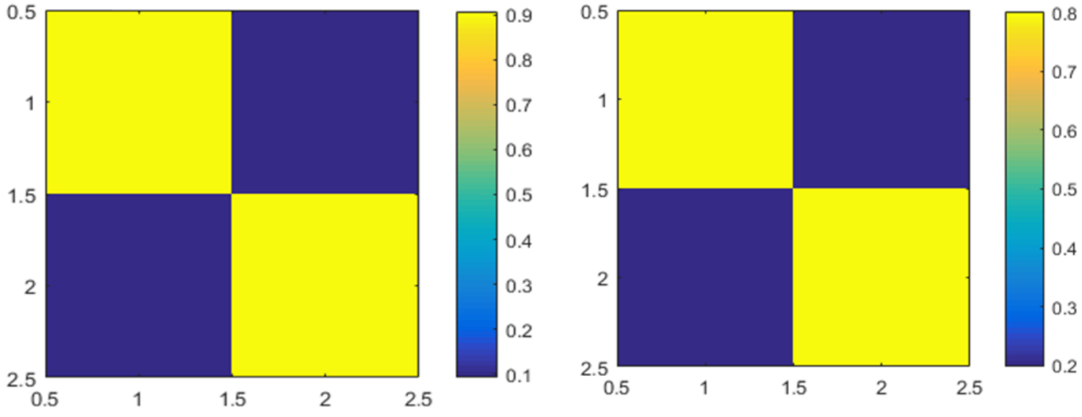


Figure 6. PCA hits.

In the study of pattern recognition and artificial intelligence, the confusion matrix is a tool that allows the visualization of the performance of an algorithm that is used in supervised learning. Each column of the matrix represents the number of predictions of each class, while each row represents the instances in the real class [11]. One of the benefits of confusion matrices is that they make it easy to see if the system is confusing two classes. This matrix is necessary in problems with large variables in which many of them are linear combinations of others, said linear combination is necessary to determine the number of variables linearly independent or in the geometric sense the dimension of space, this last characteristic is intrinsic to all physical system since the redundancy of dimensions distorts the information of the same.

If in the input data, the number of samples of different classes changes greatly the error rate of the classifier is not representative of how well the classifier performs the task. If for example there are 990 samples of class 1 and only 10 of class 2, the classifier can easily have a bias towards class 1. If the classifier classifies all samples as class 1, its accuracy will be 99%. This does not mean that it is a good classifier, because it had a 100% error in the classification of the samples of class 2. Using algorithms in MATLAB, the confusion matrices for the training data, Figure 7(a), validation, Figure 7(b), and for each of the models studied with PCA, Figure 8, is presented.

According to the Figure 6, the model that has the highest performance is the SVM model (yellow line) for the extraction of characteristics with PCA, that is, the percentage of success 85% for few characteristics (extraction of 5 features). As the characteristics increase with these two models, the success percentages increase in comparison with the Bayes Naive model, which deteriorates from 6 characteristics. With these graphs the SVM model would be chosen since the percentages of success for the training are high in comparison with the other models [12].



(a) (b)

Figure 7. Matrix of confusion (a) training; (b) validation.

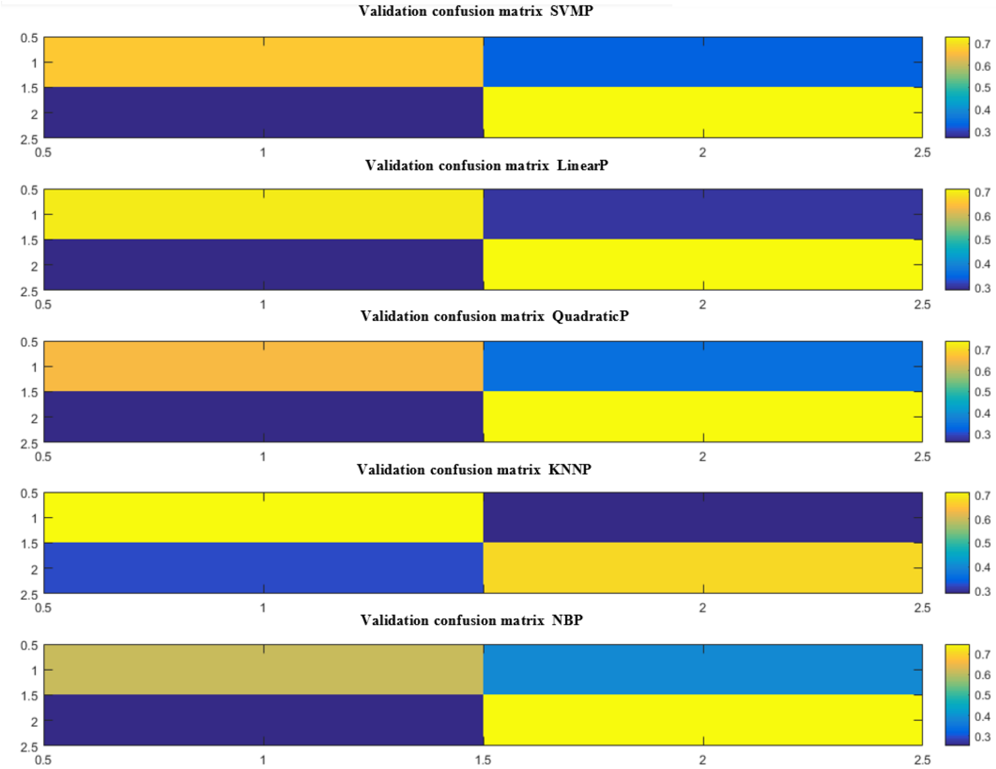


Figure 8. PCA confusion matrix.

4. Conclusions

With the use of different classification methods based on the sciences of mathematics and physics to make predictions about the behavior of a database, a comparison is made between them and the advantages and disadvantages of each are also studied. one. Among the great differences that they present, one of them is the computation times that each algorithm has, in some cases these times are very high and for this reason reductions must be made to more relevant characteristics.

The results obtained in this document allow us to classify the moments of time in which there are values of exponential moving A14 below the average to invest in the stock market (label value 0) or the moments in which it is better to wait for prices. fall to invest or trade stock selling stocks (tag value 1). These results are only informative since the variations in the stock market depend on 11 other characteristics that cannot be controlled, and the values of the market shares can overflow and never rise to carry out sales maneuvers or close transactions. used in more theoretical problems of physics which necessarily need to explain the correlation between the variables.

References

[1] García P, Sancho J L, Figueiras A 2010 Clasificación de patrones con datos faltantes: una revision *Neural Comput. Appl.* **19(2)** 263

[2] Bishop C M 1995 *Neural Networks for Pattern Recognition* (New York: Oxford University Press)

[3] Deng X, Liu Q, Deng Y, Mahadevan S 2016 An improved method to construct basic probability assignment based on the confusion matrix for classification problem *Inf. Sci.* **340** 250

[4] Rajendran S, Kaul A, Nath R, Arora A, Chauhan S 2014 Comparison of PCA and 2D-PCA on Indian faces *International Conference on Signal Propagation and Computer Technology Technical* (Ajmer: Institute of Electrical and Electronics Engineers) p 561

[5] Gao Q, Sun S 2018 Speech recognition of confusable word based on ideal distance matrix IEEE/ACIS *17th International Conference on Computer and Information Science* (Singapore: Institute of Electrical and Electronics Engineers) p 335

[6] Xiang S, Nie F, Zhang C 2008 Learning a mahalanobis distance metric for data clustering and classification *Pattern Recognition* **41(12)** 3600

[7] Jing C, Hou J 2015 Enfoques de clasificación de fallas basados en SVM y PCA para procesos industriales complicados *Neurocomputing* **167** 636

[8] Rouabhia C, Hamdaoui K, Tebbikh H 2010 Weighted matrix distance metric for face images classification *International Conference on Machine and Web Intelligence* (Algeria: Institute of Electrical and Electronics Engineers)

[9] Koyuturk M, Grama A, Ramakrishnan N 2005 Compression, clustering, and pattern discovery in very high-dimensional discrete-attribute data sets *IEEE Transactions on Knowledge and Data Engineering* **17(4)** 447

[10] Arya S, Mount D, Netanyahu N, Silverman R, Wu A 1998 An optimal algorithm for approximate nearest neighbor searching fixed dimensions *Journal of the ACM* **45(6)** 891

[11] Abello J, Korn A 2002 A system for visualizing massive multidigraph *IEEE Trans. Visual Comput. Graphics* **8(1)** 21

[12] Beringer J, Hüllermeier E 2006 Online clustering of parallel data streams *Data Knowl. Eng.* **58(2)** 180