Influential Article Review - Using Hybrid Machine Learning Algorithms in Predicting Stock Market Patters

Damian Wallis

Danika Stafford

This paper examines finance. We present insights from a highly influential paper. Here are the highlights from this paper: Big data analytic techniques associated with machine learning algorithms are playing an increasingly important role in various application fields, including stock market investment. However, few studies have focused on forecasting daily stock market returns, especially when using powerful machine learning techniques, such as deep neural networks (DNNs), to perform the analyses. DNNs employ various deep learning algorithms based on the combination of network structure, activation function, and model parameters, with their performance depending on the format of the data representation. This paper presents a comprehensive big data analytics process to predict the daily return direction of the SPDR S&P 500 ETF (ticker symbol: SPY) based on 60 financial and economic features. DNNs and traditional artificial neural networks (ANNs) are then deployed over the entire preprocessed but untransformed dataset, along with two datasets transformed via principal component analysis (PCA), to predict the daily direction of future stock market index returns. While controlling for overfitting, a pattern for the classification accuracy of the DNNs is detected and demonstrated as the number of the hidden layers increases gradually from 12 to 1000. Moreover, a set of hypothesis testing procedures are implemented on the classification, and the simulation results show that the DNNs using two PCA-represented datasets give significantly higher classification accuracy than those using the entire untransformed dataset, as well as several other hybrid machine learning algorithms. In addition, the trading strategies guided by the DNN classification process based on PCA-represented data perform slightly better than the others tested, including in a comparison against two standard benchmarks. For our overseas readers, we then present the insights from this paper in Spanish, French, Portuguese, and German.

Keywords: Daily stock return forecasting, Return direction classification, Data representation, Hybrid machine learning algorithms, Deep neural networks (DNNs), Trading strategies

SUMMARY

- A comprehensive big data analytics procedure using hybrid machine learning algorithms has been developed to forecast the daily return direction of the SPDR S&P 500 ETF . Ideally, researchers look to apply the simplest set of algorithms to the least amount of data, with both the most accurate forecasting results and the highest risk-adjusted profits being desired. We have also considered this standard for this research.

- The analytic process starts with data cleaning and preprocessing and concludes with an analysis of the forecasting and simulation results. It is also observed that as the number of DNN hidden layers increases, a pattern regarding the classification accuracy emerges, with the overfitting issue remaining under control. In addition, over three data sets with different representations, the trading strategies using the DNN classifiers perform better than the ones using the ANN classifiers in most cases. Although in general there is no significant difference among the trading strategies from the DNN classification process over the entire untransformed data set and two PCA-represented data sets, the trading strategies based on the PCA-represented data perform slightly better.
- Thus, when combined with the new results as illustrated in Tables 2, 3, 4 and 6, 7 8 it can be concluded that among the machine learning techniques considered in this study series, the PCA-DNN classifiers with the proper number of hidden layers can achieve the highest classification accuracy and result in the best trading strategy performance.
- With additional hidden layers and more complicated learning algorithms, DNNs are recognized as an important and advanced technology in the fields of computational intelligence and artificial intelligence.

HIGHLY INFLUENTIAL ARTICLE

We used the following article as a basis of our evaluation:

Zhong, X., & Enke, D. (2019). Predicting the daily return direction of the stock market using hybrid machine learning algorithms. Financial Innovation, 5(1), 1–20.

This is the link to the publisher's website:
https://jfin-swufe.springeropen.com/articles/10.1186/s40854-019-0138-0

INTRODUCTION

Big data analytic techniques developed with machine learning algorithms are gaining more attention in various application fields, including stock market investment. This is mainly because machine learning algorithms do not require any assumptions about the data and often achieve higher accuracy than econometric and statistical models; for example, artificial neural networks (ANNs), fuzzy systems, and genetic algorithms are driven by multivariate data with no required assumptions. Many of these methodologies have been applied to forecast and analyze financial variables, for instance, see Vellido, Lisboa, & Meehan (1999); Kim & Han (2000); Cao & Tay (2001); Thawornwong, Dagli, & Enke (2001); Bogullu, Enke, & Dagli (2002); Hansen & Nelson (2002); Wang (2002); Chen, Leung, & Daouk (2003); Zhang (2003); Chun & Kim (2004); Shen & Loh (2004); Thawornwong & Enke (2004); Armano, Marchesi, & Murru (2005); Enke & Thawornwong (2005); Ture & Kurt (2006); Amornwattana et al. (2007); Enke & Mehdiyev (2013); Zhong & Enke (2017a, 2017b); Huang & Kou (2014); Huang, Kou, & Peng (2017); and Nayak & Misra (2018). A comprehensive review of these studies was conducted by Atsalakis & Valavanis (2009) and Vanstone & Finnie (2009). With nonlinear, data-driven, and easy-to-generalize characteristics, multivariate analysis with ANNs has become a dominant and popular analysis tool in finance and economics. Refenes, Burgess, & Bentz (1997) and Zhang, Patuwo, & Hu (1998) review the use of using ANNs as a forecasting method in different areas of finance and investing, including financial engineering.

Recently, deep learning has emerged as a powerful machine learning technique owing to its far-reaching implications for artificial intelligence, although deep learning methods are not currently considered as an all-encompassing solution for the effective application of artificial intelligence. ANNs using different deep learning algorithms are categorized as deep neural networks (DNNs), which have been applied to many important fields, such as automatic speech recognition, image recognition, natural language processing, drug discovery and toxicology, customer relationship management, recommendation systems, and bioinformatics where they have often been shown to produce improved results for different tasks.

Moreover, it is critical for neural networks with different topologies to achieve accurate results with a deliberate selection of input variables (Lam, 2004; Hussain et al., 2007). The most influential and representative inputs can be chosen using mature dimensionality reduction technologies, such as principal component analysis (PCA), and its variants fuzzy robust principal component analysis (FRPCA) and kernel-based principal component analysis (KPCA), among others. PCA is a classical and well-known statistical linear method for extracting the most influential features from a high-dimensional data space. van der Maaten et al. (2009) compare PCA with 12 front-ranked nonlinear dimensionality reduction techniques, such as multidimensional scaling, Isomap, maximum variance unfolding, KPCA, diffusion maps, multilayer autoencoders, locally linear embedding, Laplacian eigenmaps, Hessian LLE, local tangent space analysis, locally linear coordination, and manifold charting, by applying each on self-created and natural tasks. The results show that although nonlinear techniques perform well on selected artificial data, none of them outperforms the traditional PCA using real-world data. In addition, Sorzano, Vargas, & Pascual-Montano (2014) state that among the available dimensionality reduction techniques, PCA and its versions, such as the standard PCA, robust PCA, sparse PCA, and KPCA, are still preferred for their simplicity and intuitiveness.

Few studies have focused on forecasting daily stock market returns using hybrid machine learning algorithms. Zhong & Enke (2017a) present a study of dimensionality reduction with an application to predict the daily return direction of the SPDR S&P 500 ETF (ticker symbol: SPY) using ANN classifiers. They compare various ANN models and find that among the PCA and its two popular variants, FRPCA and KPCA, PCA-based ANN classifiers are shown to be the best predictor of the ETF daily return direction over various datasets transformed using PCA (Zhong & Enke, 2017a). Also, Zhong & Enke (2017b) perform a comprehensive data mining procedure, including both cluster and classification mining, to forecast the ETF daily return direction. They show that PCA-based ANN classifiers lead to significantly higher accuracy than three different PCA-based logistic regression models, including those that have successfully used fuzzy c-means clustering. Chong, Han, & Park (2017) recently examine the advantages and drawbacks of using deep learning algorithms for stock analysis and prediction, but their study focuses on intraday stock return forecasting.

In this study, the daily return direction of the SPDR S&P 500 ETF is forecasted using a deliberately designed classification mining procedure based on hybrid machine learning algorithms. This process begins by preprocessing the raw data to deal with missing values, outliers, and mismatched samples. The ANNs and DNNs, each acting as classifiers, are then used with both the entire untransformed dataset and the PCA-represented datasets to forecast the direction of future daily market returns. The remainder of this paper discusses the details of the study and is organized as follows. The data description and preprocessing are introduced next, including the transformation of the entire data set via PCA. The architectures, network topology, and learning algorithms of the newly developed DNNs, along with the previously successful benchmark ANNs, both of which are used for return direction classification, are then discussed. The forecasting procedure of three different datasets with the DNN classifiers are then described, together with the classification results and the pattern of the classification accuracy relevant to the number of hidden layers. A standard benchmark is also compared with the PCA-based ANN classifiers results. The simulation results from trading strategies based on the DNN classifiers over the three datasets are compared to each other, and the results of the ANN-based trading strategies as compared with two benchmarks are then discussed. Finally, concluding remarks and proposed future work are provided.

CONCLUSION

A comprehensive big data analytics procedure using hybrid machine learning algorithms has been developed to forecast the daily return direction of the SPDR S&P 500 ETF (ticker symbol: SPY). Ideally, researchers look to apply the simplest set of algorithms to the least amount of data, with both the most accurate forecasting results and the highest risk-adjusted profits being desired. We have also considered this standard for this research.

The analytic process starts with data cleaning and preprocessing and concludes with an analysis of the forecasting and simulation results. The comparison of the classification and simulation results is done with statistical hypothesis tests, showing that on average, the accuracy of the DNN-based classification is significantly higher than the PCA-represented data over the entire untransformed data set. More specifically, the DNN-based classification for the PCA-represented data set with PCs = 31 achieves the highest accuracy. It is also observed that as the number of DNN hidden layers increases, a pattern regarding the classification accuracy (as compared to the ANN classifier) emerges, with the overfitting issue remaining under control. In addition, over three data sets with different representations, the trading strategies using the DNN classifiers perform better than the ones using the ANN classifiers in most cases. Although in general there is no significant difference among the trading strategies from the DNN classification process over the entire untransformed data set and two PCA-represented data sets, the trading strategies based on the PCA-represented data perform slightly better.

In previous studies (Zhong & Enke, 2017a, 2017b), the PCA-ANN classifiers are shown to give a higher prediction accuracy for the daily return direction of the SPY ETF for the next day than the FRPCA-ANN classifiers, KPCA-ANN classifiers, and logistic regression classifiers, with or without PCA/FRPCA/KPCA involved. Also, the trading strategies based on the PCA-ANN classifiers perform better than the other strategies based on the other classifiers. Moreover, when using PCA, all classification model-based trading strategies perform better than the benchmark one-month T-bill strategy; the trading strategies from the ANN classification mining procedure perform better than the benchmark buy-and-hold strategy. Thus, when combined with the new results as illustrated in Tables 2, 3, 4 and 6, 7 8 it can be concluded that among the machine learning techniques considered in this study series, the PCA-DNN classifiers with the proper number of hidden layers can achieve the highest classification accuracy and result in the best trading strategy performance.

With additional hidden layers and more complicated learning algorithms, DNNs are recognized as an important and advanced technology in the fields of computational intelligence and artificial intelligence. However, DNNs are still regarded as a black box with less clear theoretical confirmations of the learning algorithms that are used in common deep architectures, such as the stochastic gradient descent methodology. These DNN learning algorithms actually increase the computation time as a large number of hidden layers and neurons are included. This area of research needs to receive more attention and effort in the future.

APPENDIX

FIGURE 1

HISTOGRAM OF SPY CURRENT RETURN (LEFT) AND HISTOGRAM OF ADJUSTED SPY CURRENT RETURN (RIGHT)

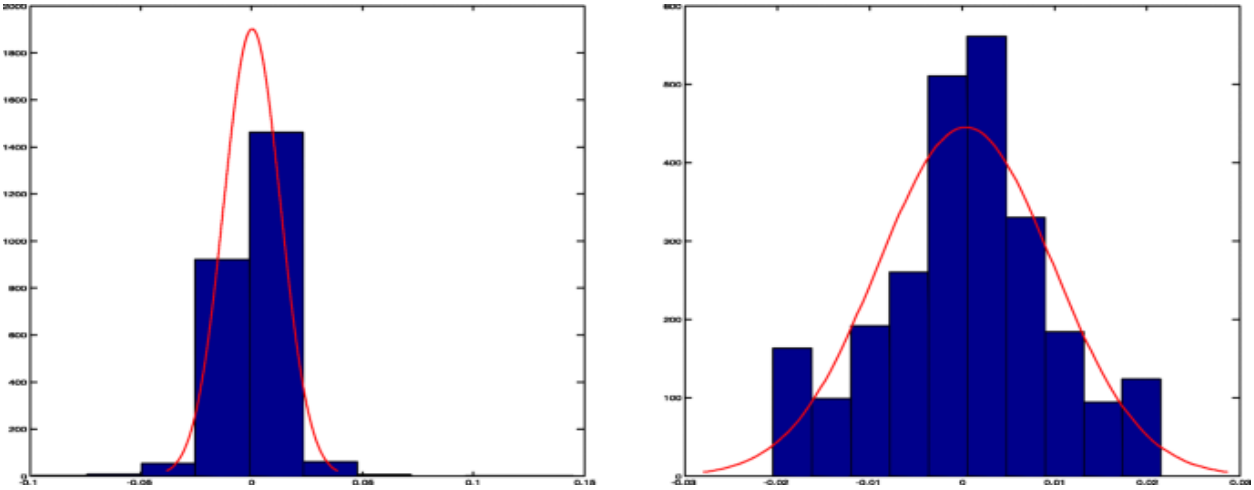

FIGURE 2
TOPOLOGY OF A MULTILAYER FEED-FORWARD NEURAL NETWORK USED FOR CLASSIFICATION

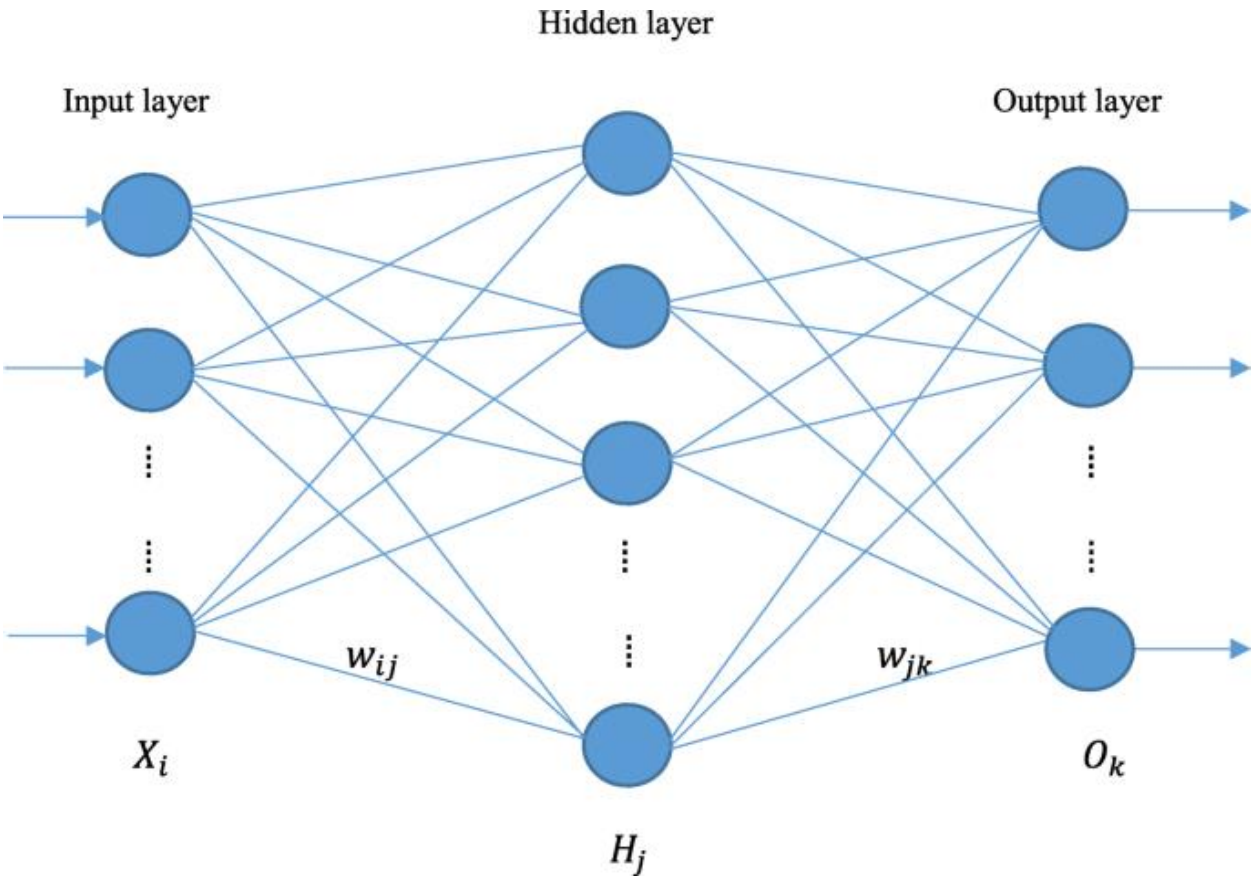

TABLE 1
THE ANN CLASSIFICATION RESULTS USING 12 TRANSFORMED DATASETS

PCs	Training	Validation	Testing	Overall
1	54.8	53.6	56.8	54.9
3	55.2	53.3	57.3	55.2
6	54.9	53.6	57.3	55
10	56.4	54.6	57.3	56.3
15	56.3	53.3	57.6	56
22	55.2	54.6	58.1	55.5
26	55.1	53.1	58.1	55.2
31	57.5	57.3	***58.1***	***57.5***
34	56.2	56	57.3	56.4
37	55	54.4	57	55.2
40	56.2	56.2	56.2	56.2
60	***57.5***	***54.1***	***58.1***	***57.1***

TABLE 2
CLASSIFICATION RESULTS WITH ANN/DNN CLASSIFIERS USING ENTIRE UNTRANSFORMED DATA

# of hidden layers	Training (MSE)	Validation (MSE)	Testing (MSE)	Total (MSE)
10	**57.3 (0.3058)**	**53.8 (0.3164)**	**57.3 (0.3124)**	**56.8 (0.3084)**
12	57.5 (0.3055)	54.1 (0.3129)	57.3 (0.3110)	56.9 (0.3074)
14	57.8 (0.3041)	53.8 (0.3127)	57.6 (0.3075)	57.2 (0.3059)
16	58.6 (0.3034)	54.9 (0.3160)	58.1 (0.3099)	57.9 (0.3063)
18	58.2 (0.3045)	53.3 (0.3143)	58.1 (0.3095)	57.5 (0.3067)
20	**59.1 (0.3052)**	**54.4 (0.3186)**	**58.4 (0.3146)**	**58.3 (0.3086)**
22	57.7 (0.3041)	54.1 (0.3169)	58.1 (0.3099)	57.2 (0.3069)
24	57.0 (0.3071)	55.7 (0.3139)	58.1 (0.3066)	57 (0.3081)
26	55.4 (0.3144)	54.9 (0.3245)	58.1 (0.3143)	55.8 (0.3159)
28	54.6 (0.3163)	54.6 (0.3175)	**58.6 (0.3074)**	55.2 (0.3151)
30	54.5 (0.3163)	53.1 (0.3232)	58.1 (0.3215)	54.8 (0.3181)
35	58.3 (0.3050)	54.9 (0.3169)	57.8 (0.3121)	57.7 (0.3079)
40	56.3 (0.3354)	53.3 (0.3584)	57.0 (0.3399)	56.0 (0.3395)
45	58.0 (0.3055)	53.8 (0.3201)	57.0 (0.3113)	57.2 (0.3085)
50	58.3 (0.3034)	53.6 (0.3252)	57.3 (0.3134)	57.4 (0.3081)
100	54.5 (0.3354)	53.3 (0.3353)	57.0 (0.3219)	54.7 (0.3334)
500	55.4 (0.3474)	53.8 (0.3570)	57.3 (0.3386)	55.5 (0.3475)
1000	57.3 (0.3383)	54.1 (0.3521)	57.3 (0.3383)	56.8 (0.3404)

TABLE 3
CLASSIFICATION RESULTS WITH ANN/DNN CLASSIFIERS USING TRANSFORMED DATA WITH PCS = 60

# of hidden layers	training (MSE)	validation (MSE)	testing (MSE)	total (MSE)
10	**58.2 (0.3062)**	**54.1 (0.3110)**	**57.8 (0.3091)**	**57.5 (0.3074)**
12	56.9 (0.3079)	53.3 (0.3137)	58.1 (0.3066)	56.6 (0.3086)
14	57.9 (0.3041)	54.6 (0.3135)	57.8 (0.3084)	57.4 (0.3062)
16	**59.4 (0.3020)**	**55.4 (0.3128)**	**59.9 (0.3056)**	**58.9 (0.3042)**
18	56.7 (0.3071)	54.6 (0.3109)	58.9 (0.3089)	56.7 (0.3080)
20	58.8 (0.3052)	54.4 (0.3109)	59.2 (0.3074)	58.2 (0.3064)
22	57.3 (0.3065)	55.4 (0.3133)	59.4 (0.3083)	57.3 (0.3078)
24	56.9 (0.3080)	54.9 (0.3099)	58.4 (0.3082)	56.8 (0.3083)
26	55.9 (0.3101)	56.0 (0.3105)	58.4 (0.3088)	56.3 (0.3099)
28	57.8 (0.3057)	56.5 (0.3105)	59.4 (0.3079)	57.9 (0.3067)
30	56.2 (0.3076)	53.6 (0.3152)	58.1 (0.3104)	56.1 (0.3092)
35	56.6 (0.3066)	56.2 (0.3134)	58.1 (0.3081)	56.8 (0.3078)
40	59.8 (0.2999)	54.9 (0.3125)	57.6 (0.3095)	58.7 (0.3032)
45	56.3 (0.3096)	54.6 (0.3163)	57.3 (0.3113)	56.2 (0.3109)
50	55.2 (0.3103)	53.6 (0.3154)	57.3 (0.3078)	55.3 (0.3107)
100	56.9 (0.3077)	53.1 (0.3205)	57.6 (0.3221)	56.4 (0.3117)
500	55.5 (0.3345)	54.9 (0.3309)	**59.9 (0.3162)**	56.1 (0.3312)
1000	58.4 (0.3240)	55.7 (0.3392)	58.1 (0.3285)	57.9 (0.3269)

TABLE 4
CLASSIFICATION RESULTS WITH ANN/DNN CLASSIFIERS USING TRANSFORMED DATA WITH PCS = 31

# of hidden layers	Training (MSE)	Validation (MSE)	Testing (MSE)	Total (MSE)
10	**56.1 (0.3067)**	**54.4 (0.3121)**	**58.9 (0.3095)**	**56.3 (0.3079)**
12	**61.6 (0.3030)**	56.8 (0.3253)	58.4 (0.3141)	**60.4 (0.3080)**
14	54.6 (0.3237)	54.9 (0.3111)	58.9 (0.3051)	55.3 (0.3190)
16	**61.0 (0.2980)**	**56.5 (0.3087)**	**59.4 (0.3084)**	**60.1 (0.3011)**
18	54.9 (0.3145)	55.4 (0.3160)	59.2 (0.3091)	55.6 (0.3139)
20	55.0 (0.3096)	56.5 (0.3083)	59.7 (0.3079)	56.0 (0.3092)
22	55.6 (0.3097)	56.8 (0.3120)	**59.9 (0.3059)**	56.4 (0.3095)
24	54.1 (0.3105)	54.1 (0.3133)	58.9 (0.3132)	54.8 (0.3113)
26	56.9 (0.3228)	54.4 (0.3191)	58.6 (0.3125)	56.8 (0.3207)
28	57.1 (0.3049)	54.9 (0.3136)	59.4 (0.3081)	57.1 (0.3067)
30	54.8 (0.3152)	54.4 (0.3142)	58.4 (0.3085)	55.2 (0.3140)
35	58.2 (0.3049)	55.4 (0.3167)	58.9 (0.3083)	57.9 (0.3072)
40	55.3 (0.3111)	54.6 (0.3163)	58.6 (0.3071)	55.7 (0.3113)
45	59.2 (0.3003)	55.7 (0.3147)	58.1 (0.3081)	58.5 (0.3036)
50	57.9 (0.3040)	54.9 (0.3140)	58.4 (0.3070)	57.5 (0.3059)
100	58.6 (0.3044)	54.4 (0.3131)	58.9 (0.3061)	58.0 (0.3060)
500	60.4 (0.3117)	55.4 (0.3436)	58.6 (0.3233)	59.4 (0.3182)
1000	57.7 (0.3237)	56.0 (0.3405)	58.9 (0.3293)	57.6 (0.3271)

TABLE 5
COMPARISON OF CLASSIFICATION RESULTS FROM DNN CLASSIFIERS FOR THREE DATA SETS

Null hypothesis	Alternative hypothesis	P-value
$\mu_{entire} = \mu_{pcs60}$	$\mu_{entire} < \mu_{pcs60}$	1.9144e-04
$\mu_{pcs60} = \mu_{pcs31}$	$\mu_{pcs60} < \mu_{pcs31}$	0.0050

TABLE 6
SIMULATION RESULTS WITH ANN/DNN CLASSIFIERS USING ENTIRE UNTRANSFORMED DATA

# of hidden layers	Mean of daily return	Std. of daily return	Sharpe ratio
10	**7.8493E-04**	**0.0077**	**0.1015**
12	7.4376E-04	0.0071	0.1051
14	8.3735E-04	0.0077	0.1090
16	8.2346E-04	0.0078	0.1056
18	1.0000E-03	0.0073	0.1411
20	7.8827E-04	0.0077	0.1030
22	8.4592E-04	0.0077	0.1103
24	8.6660E-04	0.0073	0.1187
26	8.8574E-04	0.0074	0.1196
28	8.3240E-04	0.0075	0.1112
30	8.4049E-04	0.0079	0.1071
35	8.6501E-04	0.0077	0.1119
40	7.9263E-04	0.0079	**0.1006**
45	8.2000E-04	0.0073	0.1125
50	7.7529E-04	0.0077	**0.1004**
100	8.4306E-04	0.0076	0.1110
500	7.9310E-04	0.0079	**0.1007**
1000	7.9541E-04	0.0078	0.1019

TABLE 7
SIMULATION RESULTS WITH ANN/DNN CLASSIFIERS USING TRANSFORMED DATA WITH PCS = 60

# of hidden layers	Mean of daily return	Std. of daily return	Sharpe ratio
10	**7.6471E-04**	**0.0076**	**0.1011**
12	8.7298E-04	0.0074	0.1178
14	7.0400E-04	0.0077	**0.0911**
16	9.0078E-04	0.0076	0.1181
18	9.0041E-04	0.0075	0.1202
20	9.6420E-04	0.0075	0.1294
22	9.0986E-04	0.0077	0.1188
24	7.8212E-04	0.0076	0.1036
26	9.6026E-04	0.0070	0.1375
28	9.5506E-04	0.0071	0.1354
30	9.3496E-04	0.0074	0.1271
35	7.9479E-04	0.0077	0.1035
40	5.8272E-04	0.0075	**0.0778**
45	7.0538E-04	0.0074	**0.0953**
50	5.9244E-04	0.0071	**0.0832**
100	8.3309E-04	0.0079	0.1061
500	9.3984E-04	0.0074	0.1275
1000	8.7984E-04	0.0076	0.1150

TABLE 8
SIMULATION RESULTS WITH ANN/DNN CLASSIFIERS USING TRANSFORMED DATA WITH PCS = 31

# of hidden layers	Mean of daily return	Std. of daily return	Sharpe ratio
10	**8.0339E-04**	**0.0076**	**0.1064**
12	7.4933E-04	0.0071	**0.1057**
14	9.3477E-04	0.0072	0.1292
16	9.3504E-04	0.0072	0.1294
18	9.6857E-04	0.0071	0.1359
20	8.0664E-04	0.0072	0.1115
22	9.6978E-04	0.0077	0.1267
24	5.7661E-04	0.0069	**0.0836**
26	7.7980E-04	0.0076	**0.1031**
28	8.5625E-04	0.0078	0.1099
30	8.4888E-04	0.0075	0.1127
35	8.5513E-04	0.0078	0.1093
40	8.2210E-04	0.0076	0.1081
45	7.8532E-04	0.0075	**0.1042**
50	7.1064E-04	0.0077	**0.0922**
100	8.2574E-04	0.0073	0.1126
500	8.9993E-04	0.0077	0.1169
1000	7.9599E-04	0.0076	**0.1050**

TABLE 9
COMPARISON OF SIMULATION RESULTS FROM DNN CLASSIFIERS FOR THREE DATA SETS

Null hypothesis	Alternative hypothesis	P-value
$\mu_{entire} = \mu_{pcs60}$	$\mu_{entire} \neq \mu_{pcs60}$	0.6251
$\mu_{pcs60} = \mu_{pcs31}$	$\mu_{pcs60} \neq \mu_{pcs31}$	0.8897
$\mu_{entire} = \mu_{pcs31}$	$\mu_{entire} \neq \mu_{pcs31}$	0.6635
$\mu_{entire} = \mu_{pcs60}$	$\mu_{entire} < \mu_{pcs60}$	0.3126
$\mu_{pcs60} = \mu_{pcs31}$	$\mu_{pcs60} < \mu_{pcs31}$	0.5552
$\mu_{entire} = \mu_{pcs31}$	$\mu_{entire} < \mu_{pcs31}$	0.3318

TABLE 10
THE 60 FINANCIAL AND ECONOMICAL FEATURES OF THE RAW DATA

Group	Name	Description	Source/Calculation
	Date_SPY	trading dates considered	finance.yahoo.com
	Close_SPY	closing prices of SPY on the trading days	finance.yahoo.com
SPY return in current and three previous days			
	SPYt	The return of the SPDR S&P 500 ETF (SPY) in day t.	finance.yahoo.com / (p(t) - p(t-1))/p(t-1)
	SPYt1	The return of the SPY in day t-1.	finance.yahoo.com / (p(t-1) - p(t-2))/p(t-2)
	SPYt2	The return of the SPY in day t-2.	finance.yahoo.com / (p(t-2) - p(t-3))/p(t-3)
	SPYt3	The return of the SPY in day t-3.	finance.yahoo.com / (p(t-3) - p(t-4))/p(t-4)
Relative difference in percentage of the SPY return			
	RDP5	The 5-day relative difference in percentage of the SPY.	(p(t) - p(t-5))/p(t-5) * 100
	RDP10	The 10-day relative difference in percentage of the SPY.	(p(t) - p(t-10))/p(t-10) * 100
	RDP15	The 15-day relative difference in percentage of the SPY.	(p(t) - p(t-15))/p(t-15) * 100
	RDP20	The 20-day relative difference in percentage of the SPY.	(p(t) - p(t-20))/p(t-20) * 100
Exponential moving averages of the SPY return			
	EMA10	The 10-day exponential moving average of the SPY.	p(t) * (2/(10+1)) + EMA10 (t-1) * (1-2/(10+1))
	EMA20	The 20-day exponential moving average of the SPY.	p(t) * (2/(20+1)) + EMA20 (t-1) * (1-2/(20+1))
	EMA50	The 50-day exponential moving average of the SPY.	p(t) * (2/(50+1)) + EMA50 (t-1) * (1-2/(50+1))
	EMA200	The 200-day exponential moving average of the SPY.	p(t) * (2/(200+1)) + EMA200 (t-1) * (1-2/(200+1))
T-bill rates (in day t)			
	T1	1-month T-bill rate, secondary market, business days, discount basis.	H. 15 Release - Federal Reserve Board of Governors (https://research.stlouisfed.org/fred2/series/DGS5/downloaddata)
	T3	3-month T-bill rate, secondary market, business days, discount basis.	H. 15 Release - Federal Reserve Board of Governors (https://research.stlouisfed.org/fred2/series/DGS5/downloaddata)
	T6	6-month T-bill rate, secondary market, business days, discount basis.	H. 15 Release - Federal Reserve Board of Governors (https://research.stlouisfed.org/fred2/series/DGS5/downloaddata)
	T60	5-year T-bill constant maturity rate, secondary market, business days.	H. 15 Release - Federal Reserve Board of Governors (https://research.stlouisfed.org/fred2/series/DGS5/downloaddata)
	T120	10-year T-bill constant maturity rate, secondary market, business days.	H. 15 Release - Federal Reserve Board of Governors(https://research.stlouisfed.org/fred2/series/DGS10?catbc=1&utm_expid=19978471-Srci7QpGidAURO4vg_Q.1&utm_referrer=https%3A%2F%2Fresearch.stlouisfed.org%2Ffred2%2Frelease%3Frid%3D18)

Certificate of deposit rates (in day t)			
	CD1	Average rate on 1-month neogtiable certificates of deposit (secondary market), quoted on an investment basis.	H. 15 Release - Federal Reserve Board of Governors
	CD3	Average rate on 3-month neogtiable certificates of deposit (secondary market), quoted on an investment basis.	H. 15 Release - Federal Reserve Board of Governors
	CD6	Average rate on 6-month neogtiable certificates of deposit (secondary market), quoted on an investment basis.	H. 15 Release - Federal Reserve Board of Governors
Financial and economical indicators (in day t)			
	Oil	Relative change in the price of the crude oil (Cushing, OK WTI Spot Price FOB (dollars per barrel)).	Energy Inormation Administration, http://tonto.eia.doe.gov/dnav/pet/hist/rwtcd.htm (work on cleaning the price column first using th SPY dates as control, then cal the relative change)
	Gold	Relative change in the gold price	usagold.com (use FireFox to Select All, then copy and paste to an Excel file) (the dates used by USAGOLD are not matching with the SPY prices from yahoo.finance. For example, after 06/09/2004. We still clean/make up/delete the gold prices based on the dates of SPY prices from finance.yahoo.com. Use the same procedure in the whole data set: Take the average of the two closest data with the missing one in the middle. Then delete the mismatching one, and cal the relatvie difference as before. Another example, the data in 2011, all Friday's prices were recorded as Sunday's prices, so we estimated Friday's prices with the average of Thursday and Sunday's prices. Then deleted Sunday's prices. If there are n continuous values missing, then take the average of the n available values on each side of these n missing values, use the average for all n missing values)
	CTB3M	Change in the market yield on US Treasury securities at 3-month constant maturity, quoted on investment basis.	H. 15 Release - Federal Reserve Board of Governors
	CTB6M	Change in the market yield on US Treasury securities at 6-month constant maturity, quoted on investment basis.	H. 15 Release - Federal Reserve Board of Governors
	CTB1Y	Change in the market yield on US Treasury securities at 1-year constant maturity, quoted on investment basis.	H. 15 Release - Federal Reserve Board of Governors
	CTB5Y	Change in the market yield on US Treasury securities at 5-year constant maturity, quoted on investment basis.	H. 15 Release - Federal Reserve Board of Governors
	CTB10Y	Change in the market yield on US Treasury securities at 10-year constant maturity, quoted on investment basis.	H. 15 Release - Federal Reserve Board of Governors
	AAA	Change in the Moody's yield on seasoned corporate bonds - all industries, Aaa.	H. 15 Release - Federal Reserve Board of Governors
	BAA	Change in the Moody's yield on seasoned corporate bonds - all industries, Baa.	H. 15 Release - Federal Reserve Board of Governors

The term and default spreads			
	TE1	Term spread between T120 and T1.	TE1 = T120 - T1
	TE2	Term spread between T120 and T3.	TE2 = T120 - T3
	TE3	Term spread between T120 and T6.	TE3 = T120 - T6
	TE5	Term spread between T3 and T1.	TE5 = T3 - T1
	TE6	Term spread between T6 and T1.	TE6 = T6 - T1
	DE1	Default spread between BAA and AAA.	DE1 = BAA - AAA
	DE2	Default spread between BAA and T120.	DE2 = BAA - T120
	DE4	Default spread between BAA and T6.	DE4 = BAA - T6
	DE5	Default spread between BAA and T3.	DE5 = BAA - T3
	DE6	Default spread between BAA and T1.	DE6 = BAA - T1
	DE7	Default spread between CD6 and T6.	DE7 = CD6 - T6
Exchange rate between USD and four other currencies (in day t)			
	USD_Y	Relative change in the exchange rate between US dollar and Japanese yen.	http://www.investing.com/currencies/usd-jpy-historical-data
	USD_GBP	Relative change in the exchange rate between US dollar and British pound.	http://www.investing.com/currencies/gbp-usd-historical-data (then, take the opposites to the changes)
	USD_CAD	Relative change in the exchange rate between US dollar and Canadian dollar.	http://www.investing.com/currencies/usd-cad-historical-data
	USD_CNY	Relative change in the exchange rate between US dollar and Chinese Yuan (Renminbi).	http://www.investing.com/currencies/usd-cny-historical-data
The return of the other seven world major indices (in day t)			
	HSI	Hang Seng index return in day t.	finance.yahoo.com
	SSE Composite	Shang Hai Stock Exchange Composite index return in day t.	finance.yahoo.com
	FCHI	CAC 40 index return in day t.	finance.yahoo.com
	FTSE	FTSE 100 index return in day t.	finance.yahoo.com
	GDAXI	DAX index return in day t.	finance.yahoo.com
	DJI	Dow Jones Industrial Average index return in day t.	finance.yahoo.com(no download function for this one); measuringworth.com/datasets/DJA/result.php
	IXIC	NASDAQ Composite index return in day t.	finance.yahoo.com
SPY trading volume (in day t)	V	Relative change in the trading volume of S&P 500 index (SPY)	finance.yahoo.com

The return of the eight big companies in S&P 500 (in day t)			
	AAPL	Apple Inc stock return in day t.	finance.yahoo.com
	MSFT	Microsoft stock return in day t.	finance.yahoo.com
	XOM	Exxon Mobil stock return in day t.	finance.yahoo.com
	GE	General Electric stock return in day t.	finance.yahoo.com
	JNJ	Johnson and Johnson stock return in day t.	finance.yahoo.com
	WFC	Wells Fargo stock return in day t.	finance.yahoo.com
	AMZN	Amazon.com Inc stock return in day t.	finance.yahoo.com
	JPM	JPMorgan Chase & Co stock return in day t.	finance.yahoo.com

REFERENCES

Aizenberg I, Aizenberg NN, Vandewalle JPL (2000) Multi-Valued and Universal Binary Neurons: Theory, Learning and Applications. Springer Science & Business Media, Boston

Amornwattana S, Enke D, Dagli C (2007) A hybrid options pricing model using a neural network for estimating volatility. Int J Gen Syst 36(5):558–573

Armano G, Marchesi M, Murru A (2005) A hybrid genetic-neural architecture for stock indexes forecasting. Inf Sci 170(1):3–33

Atsalakis GS, Valavanis KP (2009) Surveying stock market forecasting techniques – part II: soft computing methods. Expert Syst Appl 36(3):5941–5950

Bogullu VK, Enke D, Dagli C (2002) Using neural networks and technical indicators for generating stock trading signals. Intell Eng Syst Art Neural Networks, Am Soc Mechanical Eng 12:721–726

Cao L, Tay F (2001) Financial forecasting using vector machines. Neural Comput & Applic 10:184–192

Chen AS, Leung MT, Daouk H (2003) Application of neural networks to an emerging financial market: forecasting and trading the Taiwan stock index. Comput Oper Res 30(6):901–923

Chiang WC, Enke D, Wu T, Wang R (2016) An adaptive stock index trading decision support system. Expert Syst Appl 59:195–207

Chong E, Han C, Park FC (2017) Deep learning networks for stock market analysis and prediction: methodology, data representations, and case studies. Expert Syst Appl 83:187–205

Chun SH, Kim SH (2004) Data mining for financial prediction and trading: application to single and multiple markets. Expert Syst Appl 26(2):131–139

Dechter R (1986) Learning while searching in constraint-satisfaction problems. AAAI-86 Proceedings, Palo Alto, pp 178–183

Enke D, Mehdiyev N (2013) Stock market prediction using a combination of stepwise regression analysis, differential evolution-based fuzzy clustering, and a fuzzy inference neural network. Intell Autom Soft Comput 19(4):636–648

Enke D, Thawornwong S (2005) The use of data mining and neural networks for forecasting stock market returns. Expert Syst Appl 29(4):927–940

Hansen JV, Nelson RD (2002) Data mining of time series using stacked generalizers. Neurocomputing 43(1–4):173–184

Huang Y, Kou G (2014) A kernel entropy manifold learning approach for financial data analysis. Decis Support Syst 64:31–42

Huang Y, Kou G, Peng Y (2017) Nonlinear manifold learning for early warning in financial markets. Eur J Oper Res 258(2):692–702
Hussain AJ, Knowles A, Lisboa PJG, El-Deredy W (2007) Financial time series prediction using polynomial pipelined neural networks. Expert Syst Appl 35:1186–1199
Ivakhnenko AG (1973) Cybernetic predicting devices. CCM Information Corporation, Amsterdam
Jolliffe T (1986) Principal component analysis. Springer-Verlag, New York
Kim KJ, Han I (2000) Genetic algorithms approach to feature discretization in artificial neural networks for the predication of stock price index. Expert Syst Appl 19(2):125–132
Kim YM, Enke D (2016) Developing a rule change trading system for the futures market using rough set analysis. Expert Syst Appl 59:165–173
Lam M (2004) Neural network techniques for financial performance prediction: integrating fundamental and technical analysis. Decis Support Syst 37:567–581
Navidi W (2011) Statistics for engineers and scientists, 3rd edn. McGraw-Hill, New York
Nayak SC, Misra BB (2018) Estimating stock closing indices using a GA-weighted condensed polynomial neural network. Financ Innov 4(21):1–22
Niaki STA, Hoseinzade S (2013) Forecasting S&P 500 index using artificial neural networks and design of experiments. J Indust Eng Int 9(1):1–9
Refenes APN, Burgess AN, Bentz Y (1997) Neural networks in financial engineering: a study in methodology. IEEE Trans Neural Netw 8(6):1222–1267
Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533–536
Shen L, Loh HT (2004) Applying rough sets to market timing decisions. Decis Support Syst 37(4):583–597
Sorzano, C. O. S., Vargas, J., & Pascual-Montano, A. (2014). A survey of dimensionality reduction techniques. arXiv: 1403.2877v1 [stat.ML]
Thawornwong S, Dagli C, Enke D (2001) Using neural networks and technical analysis indicators for predicting stock trends. Intelligent Engineering Systems through Artificial Neural Networks. Am Soc Mech Eng 11:739–744
Thawornwong S, Enke D (2004) The adaptive selection of financial and economic variables for use with artificial neural networks. Neurocomputing 56:205–232
Ture M, Kurt I (2006) Comparison of four different time series methods to forecast hepatitis a virus infection. Expert Syst Appl 31(1):41–46
van der Maaten LJ, Postma EO, van den Herik HJ (2009) Dimensionality reduction: a comparative review. J Mach Learn Res 10(1–41):66–71
Vanstone B, Finnie G (2009) An empirical methodology for developing stock market trading systems using artificial neural networks. Expert Syst Appl 36(3):6668–6680
Vellido A, Lisboa PJG, Meehan K (1999) Segmentation of the on-line shopping market using neural networks. Expert Syst Appl 17(4):303–314
Wang YF (2002) Predicting stock price using fuzzy grey prediction system. Expert Syst Appl 22(1):33–39
Zhang G (2003) Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50:159–175
Zhang G, Patuwo BE, Hu MY (1998) Forecasting with artificial neural networks: the state of the art. Int J Forecast 14(1):35–62
Zhong X, Enke D (2017a) Forecasting daily stock market return using dimensionality reduction. Expert Syst Appl 67:126–139
Zhong X, Enke D (2017b) A comprehensive cluster and classification mining procedure for daily stock market return forecasting. Neurocomputing 267:152–168

TRANSLATED VERSION: SPANISH

Below is a rough translation of the insights presented above. This was done to give a general understanding of the ideas presented in the paper. Please excuse any grammatical mistakes and do not hold the original authors responsible for these mistakes.

VERSION TRADUCIDA: ESPAÑOL

A continuación se muestra una traducción aproximada de las ideas presentadas anteriormente. Esto se hizo para dar una comprensión general de las ideas presentadas en el documento. Por favor, disculpe cualquier error gramatical y no responsabilite a los autores originales de estos errores.

INTRODUCCIÓN

Las técnicas analíticas de Big Data desarrolladas con algoritmos de aprendizaje automático están ganando más atención en varios campos de aplicación, incluida la inversión en el mercado de valores. Esto se debe principalmente a que los algoritmos de aprendizaje automático no requieren ninguna suposición sobre los datos y a menudo logran una mayor precisión que los modelos econométricos y estadísticos; por ejemplo, las redes neuronales artificiales (ANN), los sistemas difusos y los algoritmos genéticos son impulsados por datos multivariados sin suposiciones requeridas. Muchas de estas metodologías se han aplicado para pronosticar y analizar variables financieras, por ejemplo, véase Vellido, Lisboa y Meehan (1999); Kim & Han (2000); Cao & Tay (2001); Thawornwong, Dagli y Enke (2001); Bogullu, Enke y Dagli (2002); Hansen & Nelson (2002); Wang (2002); Chen, Leung y Daouk (2003); Zhang (2003); Chun & Kim (2004); Shen & Loh (2004); Thawornwong & Enke (2004); Armano, Marchesi y Murru (2005); Enke & Thawornwong (2005); Ture & Kurt (2006); Amornwattana et al. (2007); Enke & Mehdiyev (2013); Zhong & Enke (2017a, 2017b); Huang & Kou (2014); Huang, Kou y Peng (2017); y Nayak & Misra (2018). Atsalakis & Valavanis (2009) y Vanstone & Finnie (2009) llevaron a cabo una revisión exhaustiva de estos estudios. Con características no lineales, basadas en datos y fáciles de generalizar, el análisis multivariado con ANN se ha convertido en una herramienta de análisis dominante y popular en finanzas y economía. Refenes, Burgess, & Bentz (1997) y Zhang, Patuwo, & Hu (1998) revisan el uso del uso de ANN como método de predicción en diferentes áreas de finanzas e inversión, incluida la ingeniería financiera.

Recientemente, el aprendizaje profundo ha surgido como una poderosa técnica de aprendizaje automático debido a sus implicaciones de gran alcance para la inteligencia artificial, aunque los métodos de aprendizaje profundo no se consideran actualmente como una solución integral para la aplicación efectiva de la inteligencia artificial. Las ANN que utilizan diferentes algoritmos de aprendizaje profundo se clasifican como redes neuronales profundas (DNN), que se han aplicado a muchos campos importantes, como el reconocimiento automático de voz, el reconocimiento de imágenes, el procesamiento del lenguaje natural, el descubrimiento y la toxicología de medicamentos, la gestión de relaciones con los clientes, los sistemas de recomendación y la bioinformática, donde a menudo se ha demostrado que producen resultados mejorados para diferentes tareas.

Además, es fundamental que las redes neuronales con diferentes topologías logren resultados precisos con una selección deliberada de variables de entrada (Lam, 2004; Hussain et al., 2007). Las entradas más influyentes y representativas se pueden elegir utilizando tecnologías de reducción de dimensionalidad maduras, como el análisis de componentes principales (PCA), y sus variantes de análisis de componentes principales robustos y difusos (FRPCA) y análisis de componentes principales basados en kernel (KPCA), entre otros. PCA es un método lineal estadístico clásico y bien conocido para extraer las características más influyentes de un espacio de datos de alta dimensión. Van der Maaten et al. (2009) comparar PCA con 12 técnicas de reducción de dimensionalidad no lineal clasificadas frontalmente, tales como escalado multidimensional, Isomap, desviación máxima desplegada, KPCA, mapas de difusión, autocodificadores multicapa, incrustación lineal local, eigenmaps laplacianos, LLE hessiano, análisis de espacio tangente local, coordinación localmente lineal y múltiples gráficos, aplicando cada uno en tareas auto-creadas y

naturales. Los resultados muestran que aunque las técnicas no lineales funcionan bien en los datos artificiales seleccionados, ninguna de ellas supera al PCA tradicional utilizando datos del mundo real. Además, Sorzano, Vargas, & Pascual-Montano (2014) afirman que entre las técnicas de reducción de dimensionalidad disponibles, PCA y sus versiones, como el ESTÁNDAR PCA, PCA robusto, PCA disperso, y KPCA, todavía son preferidos por su simplicidad e intuición.

Pocos estudios se han centrado en la previsión de rendimientos diarios del mercado de valores utilizando algoritmos híbridos de aprendizaje automático. Zhong & Enke (2017a) presenta un estudio de reducción de dimensionalidad con una aplicación para predecir la dirección de retorno diaria del ETF SPDR S&P 500 (símbolo de ticker: SPY) utilizando clasificadores ANN. Comparan varios modelos de ANN y encuentran que entre el PCA y sus dos variantes populares, FRPCA y KPCA, los clasificadores ANN basados en PCA han demostrado ser el mejor predictor de la dirección de retorno diaria del ETF sobre varios conjuntos de datos transformados utilizando PCA (Zhong & Enke, 2017a). Además, Zhong & Enke (2017b) realiza un procedimiento integral de minería de datos, que incluye minería de clústeres y de clasificación, para pronosticar la dirección de retorno diaria del ETF. Muestran que los clasificadores ANN basados en PCA conducen a una precisión significativamente mayor que tres modelos de regresión logística basados en PCA diferentes, incluidos aquellos que han utilizado con éxito la agrupación en clústeres de medios c difusos. Chong, Han, & Park (2017) examinan recientemente las ventajas e inconvenientes del uso de algoritmos de aprendizaje profundo para el análisis y la predicción de acciones, pero su estudio se centra en la previsión de retorno de acciones intradía.

En este estudio, se pronostica la dirección de retorno diaria del ETF SPDR S&P 500 utilizando un procedimiento de minería de clasificación diseñado deliberadamente basado en algoritmos híbridos de aprendizaje automático. Este proceso comienza preprocesando los datos sin procesar para tratar los valores que faltan, los valores atípicos y las muestras no coincidentes. Los ANN y las DNN, cada uno de los que actúan como clasificadores, se utilizan con todo el conjunto de datos notransformado y los conjuntos de datos representados por PCA para pronosticar la dirección de los rendimientos futuros del mercado diario. El resto de este documento analiza los detalles del estudio y se organiza de la siguiente manera. La descripción de los datos y el preprocesamiento se introducen a continuación, incluida la transformación de todo el conjunto de datos a través de PCA. Las arquitecturas, la topología de red y los algoritmos de aprendizaje de las DNN recién desarrolladas, junto con los ANN de referencia previamente exitosos, que se utilizan para la clasificación de la dirección de retorno, se discuten a continuación. A continuación, se describe el procedimiento de previsión de tres datasets diferentes con los clasificadores DNN, junto con los resultados de clasificación y el patrón de la precisión de clasificación relevante para el número de capas ocultas. También se compara un punto de referencia estándar con los resultados de los clasificadores ANN basados en PCA. Los resultados de la simulación de las estrategias comerciales basadas en los clasificadores DNN sobre los tres conjuntos de datos se comparan entre sí, y los resultados de las estrategias de trading basadas en ANN en comparación con dos puntos de referencia se discuten. Por último, se proporcionan observaciones finales y la labor futura propuesta.

CONCLUSIÓN

Se ha desarrollado un completo procedimiento de análisis de big data utilizando algoritmos híbridos de aprendizaje automático para pronosticar la dirección de retorno diaria del ETF SPDR S&P 500 (símbolo de ticker: SPY). Idealmente, los investigadores buscan aplicar el conjunto más simple de algoritmos a la menor cantidad de datos, con los resultados de previsión más precisos y los beneficios más altos ajustados al riesgo que se desean. También hemos considerado este estándar para esta investigación.

El proceso analítico comienza con la limpieza y el preprocesamiento de datos y concluye con un análisis de los resultados de la previsión y la simulación. La comparación de los resultados de clasificación y simulación se realiza con pruebas de hipótesis estadísticas, lo que muestra que, en promedio, la precisión de la clasificación basada en DNN es significativamente mayor que los datos representados por PCA en todo el conjunto de datos notransformados. Más específicamente, la clasificación basada en DNN para el conjunto de datos representado por PCA con pcs 31 logra la mayor precisión. También se observa que a

medida que aumenta el número de capas ocultas DNN, surge un patrón con respecto a la precisión de clasificación (en comparación con el clasificador ANN), con el problema de sobreajuste que permanece bajo control. Además, en tres conjuntos de datos con diferentes representaciones, las estrategias comerciales que utilizan los clasificadores DNN funcionan mejor que los que utilizan los clasificadores ANN en la mayoría de los casos. Aunque en general no hay ninguna diferencia significativa entre las estrategias de trading del proceso de clasificación DNN en todo el conjunto de datos notransformados y dos conjuntos de datos representados por PCA, las estrategias de trading basadas en los datos representados por PCA funcionan ligeramente mejor.

En estudios anteriores (Zhong & Enke, 2017a, 2017b), se ha demostrado que los clasificadores PCA-ANN ofrecen una mayor precisión de predicción para la dirección de retorno diaria del ETF SPY para el día siguiente que los clasificadores FRPCA-ANN, los clasificadores KPCA-ANN y los clasificadores de regresión logística, con o sin PCA/FRCAP/KPCA involucrados. Además, las estrategias de trading basadas en los clasificadores PCA-ANN funcionan mejor que las otras estrategias basadas en los otros clasificadores. Además, cuando se utiliza el PCA, todas las estrategias comerciales basadas en modelos de clasificación funcionan mejor que la estrategia de referencia de factura T de un mes; las estrategias comerciales del procedimiento de minería de clasificación ANN funcionan mejor que la estrategia de compra y retención de referencia. Por lo tanto, cuando se combina con los nuevos resultados como se ilustra en las Tablas 2, 3, 4 y 6, 7 8 se puede concluir que entre las técnicas de aprendizaje automático consideradas en esta serie de estudio, los clasificadores PCA-DNN con el número adecuado de capas ocultas pueden lograr la mayor precisión de clasificación y dar lugar al mejor rendimiento de la estrategia comercial.

Con capas ocultas adicionales y algoritmos de aprendizaje más complicados, las DNN son reconocidas como una tecnología importante y avanzada en los campos de la inteligencia computacional y la inteligencia artificial. Sin embargo, las DNN todavía se consideran como una caja negra con confirmaciones teóricas menos claras de los algoritmos de aprendizaje que se utilizan en arquitecturas profundas comunes, como la metodología de descenso de gradiente estocástico. Estos algoritmos de aprendizaje DNN realmente aumentan el tiempo de cálculo como un gran número de capas ocultas y neuronas se incluyen. Esta área de investigación necesita recibir más atención y esfuerzo en el futuro.

TRANSLATED VERSION: FRENCH

Below is a rough translation of the insights presented above. This was done to give a general understanding of the ideas presented in the paper. Please excuse any grammatical mistakes and do not hold the original authors responsible for these mistakes.

VERSION TRADUITE: FRANÇAIS

Voici une traduction approximative des idées présentées ci-dessus. Cela a été fait pour donner une compréhension générale des idées présentées dans le document. Veuillez excuser toutes les erreurs grammaticales et ne pas tenir les auteurs originaux responsables de ces erreurs.

INTRODUCTION

Les techniques d'analyse du Big Data développées avec des algorithmes d'apprentissage automatique gagnent en attention dans divers domaines d'application, y compris l'investissement boursier. C'est principalement parce que les algorithmes d'apprentissage automatique ne nécessitent aucune hypothèse sur les données et atteignent souvent une plus grande précision que les modèles économétriques et statistiques; par exemple, les réseaux neuronaux artificiels (ANN), les systèmes flous et les algorithmes génétiques sont entraînés par des données multivariées sans hypothèses requises. Bon nombre de ces méthodologies ont été appliquées pour prévoir et analyser les variables financières, par exemple, voir Vellido, Lisboa et Meehan (1999); Kim et Han (2000); Cao et Tay (2001); Thawornwong, Dagli et Enke (2001); Bogullu, Enke et Dagli (2002); Hansen et Nelson (2002); Wang (2002); Chen, Leung et Daouk (2003); Zhang (2003); Chun

et Kim (2004); Shen et Loh (2004); Thawornwong et Enke (2004); Armano, Marchesi et Murru (2005); Enke et Thawornwong (2005); Ture et Kurt (2006); Amornwattana et coll. (2007); Enke et Mehdiyev (2013); Zhong & Enke (2017a, 2017b); Huang et Kou (2014); Huang, Kou et Peng (2017); et Nayak & Misra (2018). Un examen complet de ces études a été effectué par Atsalakis et Valavanis (2009) et Vanstone & Finnie (2009). Avec des caractéristiques non lignenelles, axées sur les données et faciles à généraliser, l'analyse multivariée avec les ANN est devenue un outil d'analyse dominant et populaire en finance et en économie. Refenes, Burgess et Bentz (1997) et Zhang, Patuwo et Hu (1998) examinent l'utilisation des ANN comme méthode de prévision dans différents domaines de la finance et de l'investissement, y compris l'ingénierie financière.

Récemment, l'apprentissage profond est apparu comme une technique puissante d'apprentissage automatique en raison de ses implications profondes pour l'intelligence artificielle, bien que les méthodes d'apprentissage profond ne soient pas actuellement considérées comme une solution globale pour l'application efficace de l'intelligence artificielle. Les ANN utilisant différents algorithmes d'apprentissage profond sont classés comme des réseaux neuronaux profonds (DNN), qui ont été appliqués à de nombreux domaines importants, tels que la reconnaissance vocale automatique, la reconnaissance d'image, le traitement du langage naturel, la découverte et la toxicologie des médicaments, la gestion de la relation client, les systèmes de recommandation et la bioinformatique où il a souvent été démontré qu'ils produisent de meilleurs résultats pour différentes tâches.

En outre, il est essentiel pour les réseaux neuronaux avec différentes topologies d'obtenir des résultats précis avec une sélection délibérée de variables d'entrée (Lam, 2004; Hussain et coll., 2007). Les intrants les plus influents et les plus représentatifs peuvent être choisis à l'aide de technologies matures de réduction de la dimensionnalité, telles que l'analyse des composants principaux (PCA), et ses variantes floues robuste analyse des composants principaux (FRPCA) et l'analyse des composants principaux basés sur le noyau (KPCA), entre autres. PCA est une méthode linéaire statistique classique et bien connue pour extraire les caractéristiques les plus influentes d'un espace de données de haute dimension. Van der Maaten et coll. (2009) comparent l'apc à 12 techniques de réduction de la dimensionnalité non linéaire classées à l'avant, telles que l'échelle multidimensionnelle, l'isomap, le déroulement maximal de la variance, le KPCA, les cartes de diffusion, les autoencodeurs multicoucouches, l'intégration linéaire locale, les eigenmaps laplaciens, le LLE de Hesse, l'analyse locale de l'espace tangente, la coordination linéaire locale et la cartographie multiple, en appliquant chacune sur des tâches auto-créées et naturelles. Les résultats montrent que bien que les techniques non lignear fonctionnent bien sur certaines données artificielles, aucune d'entre elles ne surpasse le PCA traditionnel à l'aide de données réelles. En outre, Sorzano, Vargas et Pascual-Montano (2014) affirment que parmi les techniques disponibles de réduction de la dimensionnalité, PCA et ses versions, telles que le PCA standard, pca robuste, PCA clairsemé, et KPCA, sont toujours préférés pour leur simplicité et leur intuition.

Peu d'études se sont concentrées sur la prévision des rendements boursiers quotidiens à l'aide d'algorithmes hybrides d'apprentissage automatique. Zhong & Enke (2017a) présente une étude de réduction de la dimensionnalité avec une application pour prédire la direction de retour quotidien de l'etf SPDR S&P 500 (symbole ticker: SPY) en utilisant des classificateurs ANN. Ils comparent divers modèles ANN et constatent que parmi le PCA et ses deux variantes populaires, FRPCA et KPCA, les classificateurs ANN basés sur pca sont montrés pour être le meilleur prédicteur de la direction de retour quotidien etf sur divers ensembles de données transformés à l'aide de PCA (Zhong & Enke, 2017a). De plus, Zhong & Enke (2017b) effectue une procédure complète d'exploration de données, y compris l'exploration de grappes et de classifications, afin de prévoir la direction du rendement quotidien de l'etf. Ils montrent que les classificateurs ANN basés sur pca mènent à une précision significativement plus élevée que trois modèles de régression logistique basés sur pca différents, y compris ceux qui ont utilisé avec succès le clustering flou de c-moyens. Chong, Han et Park (2017) examinent récemment les avantages et les inconvénients de l'utilisation d'algorithmes d'apprentissage profond pour l'analyse et la prévision des stocks, mais leur étude se concentre sur les prévisions intrajournalnières de rendement des stocks.

Dans cette étude, la direction de retour quotidien de l'etf SPDR S&P 500 est prévue à l'aide d'une procédure d'extraction de classification délibérément conçue basée sur des algorithmes hybrides

d'apprentissage automatique. Ce processus commence par le prétraitement des données brutes pour traiter les valeurs manquantes, les valeurs aberrantes et les échantillons dépareillés. Les ANN et les DNN, qui agissent chacun en tant que classificateurs, sont ensuite utilisés avec l'ensemble de l'ensemble des données non traduites et les ensembles de données représentés par pca pour prévoir l'orientation des rendements quotidiens futurs du marché. Le reste de cet article traite des détails de l'étude et est organisé comme suit. La description des données et le prétraitement sont ensuite introduits, y compris la transformation de l'ensemble des données via PCA. Les architectures, la topologie du réseau et les algorithmes d'apprentissage des NND nouvellement développés, ainsi que les ANN de référence précédemment couronnés de succès, qui sont tous deux utilisés pour la classification des directions de retour, sont ensuite discutés. La procédure de prévision de trois ensembles de données différents avec les classificateurs DNN est ensuite décrite, ainsi que les résultats de classification et le modèle de l'exactitude de classification pertinent au nombre de couches cachées. Un indice de référence standard est également comparé aux résultats des classificateurs ANN basés sur pca. Les résultats de simulation des stratégies de trading basées sur les classificateurs DNN sur les trois ensembles de données sont comparés les uns aux autres, et les résultats des stratégies de trading basées sur ANN par rapport à deux benchmarks sont ensuite discutés. Enfin, des remarques finales et des travaux futurs proposés sont fournis.

CONCLUSION

Une procédure complète d'analyse du Big Data à l'aide d'algorithmes hybrides d'apprentissage automatique a été développée pour prévoir la direction de retour quotidien du SPDR S&P 500 ETF (symbole ticker : SPY). Idéalement, les chercheurs semblent appliquer l'ensemble d'algorithmes le plus simple à la moindre quantité de données, avec à la fois les résultats de prévision les plus précis et les bénéfices ajustés au risque les plus élevés étant souhaités. Nous avons également examiné cette norme pour cette recherche.

Le processus analytique commence par le nettoyage et le prétraitement des données et se termine par une analyse des résultats de prévision et de simulation. La comparaison des résultats de classification et de simulation se fait avec des tests d'hypothèses statistiques, montrant qu'en moyenne, l'exactitude de la classification basée sur le DNN est significativement plus élevée que les données représentées par pca sur l'ensemble des données non traduites. Plus précisément, la classification basée sur le DNN pour l'ensemble de données représentés par PCA avec PC = 31 atteint la plus haute précision. On observe également qu'à mesure que le nombre de couches cachées DNN augmente, un modèle concernant l'exactitude de classification (par rapport au classificateur ANN) apparaît, la question du surafaillage restant sous contrôle. En outre, plus de trois ensembles de données avec des représentations différentes, les stratégies de trading utilisant les classificateurs DNN sont plus performantes que celles utilisant les classificateurs ANN dans la plupart des cas. Bien qu'en général il n'y ait pas de différence significative entre les stratégies de négociation du processus de classification DNN sur l'ensemble de l'ensemble de données non traduites et deux ensembles de données représentés par pca, les stratégies de négociation basées sur les données représentées par pca fonctionnent légèrement mieux.

Dans des études antérieures (Zhong & Enke, 2017a, 2017b), les classificateurs PCA-ANN donnent une précision de prédiction plus élevée pour la direction de retour quotidien de l'etf SPY pour le lendemain que les classificateurs FRPCA-ANN, les classificateurs KPCA-ANN et les classificateurs de régression logistique, avec ou sans PCA/FRPCA/KPCA impliqués. En outre, les stratégies de trading basées sur les classificateurs PCA-ANN sont plus performantes que les autres stratégies basées sur les autres classificateurs. En outre, lors de l'utilisation de PCA, toutes les stratégies de négociation basées sur des modèles de classification obtiennent de meilleurs résultats que la stratégie de référence d'un mois en matière de facture de T; les stratégies de négociation de la procédure d'extraction de classification ANN sont plus performantes que la stratégie d'achat et de prise de référence. Ainsi, lorsqu'ils sont combinés avec les nouveaux résultats illustrés dans les tableaux 2, 3, 4 et 6, 7 8, on peut conclure que parmi les techniques d'apprentissage automatique considérées dans cette série d'études, les classificateurs PCA-DNN avec le

nombre approprié de couches cachées peuvent atteindre la plus haute précision de classification et obtenir les meilleures performances de stratégie de négociation.

Avec des couches cachées supplémentaires et des algorithmes d'apprentissage plus compliqués, les DNN sont reconnus comme une technologie importante et avancée dans les domaines de l'intelligence computationnelle et de l'intelligence artificielle. Cependant, les DNN sont toujours considérés comme une boîte noire avec des confirmations théoriques moins claires des algorithmes d'apprentissage qui sont utilisés dans les architectures profondes communes, telles que la méthodologie de descente de gradient stochastique. Ces algorithmes d'apprentissage DNN augmentent en fait le temps de calcul car un grand nombre de couches cachées et de neurones sont inclus. Ce domaine de recherche doit recevoir plus d'attention et d'efforts à l'avenir.

TRANSLATED VERSION: GERMAN

Below is a rough translation of the insights presented above. This was done to give a general understanding of the ideas presented in the paper. Please excuse any grammatical mistakes and do not hold the original authors responsible for these mistakes.

ÜBERSETZTE VERSION: DEUTSCH

Hier ist eine ungefähre Übersetzung der oben vorgestellten Ideen. Dies wurde getan, um ein allgemeines Verständnis der in dem Dokument vorgestellten Ideen zu vermitteln. Bitte entschuldigen Sie alle grammatikalischen Fehler und machen Sie die ursprünglichen Autoren nicht für diese Fehler verantwortlich.

EINLEITUNG

Big-Data-Analysetechniken, die mit Machine Learning-Algorithmen entwickelt wurden, gewinnen in verschiedenen Anwendungsbereichen, einschließlich Börseninvestitionen, mehr Aufmerksamkeit. Dies liegt vor allem daran, dass Algorithmen für maschinelles Lernen keine Annahmen über die Daten erfordern und oft eine höhere Genauigkeit als ökonometrisch-statistische Modelle erreichen; Beispielsweise werden künstliche neuronale Netzwerke (anns), Fuzzy-Systeme und genetische Algorithmen von multivariaten Daten ohne erforderliche Annahmen angetrieben. Viele dieser Methoden wurden angewendet, um Finanzvariablen vorherzusagen und zu analysieren, z. B. Siehe Vellido, Lisboa und Meehan (1999); Kim & Han (2000); Cao & Tay (2001); Thawornwong, Dagli, & Enke (2001); Bogullu, Enke, & Dagli (2002); Hansen & Nelson (2002); Wang (2002); Chen, Leung, & Daouk (2003); Zhang (2003); 2004: Chun & Kim Shen & Loh (2004); Thawornwong & Enke (2004); Armano, Marchesi,& Murru (2005); Enke & Thawornwong (2005); Ture & Kurt (2006); Amornwattana et al. (2007); Enke & Mehdiyev (2013); Zhong & Enke (2017a, 2017b); Huang & Kou (2014); Huang, Kou, & Peng (2017); und Nayak & Misra (2018). Atsalakis & Valavanis (2009) und Vanstone & Finnie (2009) führten eine umfassende Überprüfung dieser Studien durch. Mit nichtlinearen, datengesteuerten und einfach zu verallgemeinernden Merkmalen ist die multivariate Analyse mit anns zu einem dominanten und beliebten Analysetool in Den Finanz- und Wirtschaftswissenschaften geworden. Refenes, Burgess, & Bentz (1997) und Zhang, Patuwo, & Hu (1998) überprüfen die Verwendung von anns als Prognosemethode in verschiedenen Bereichen der Finanzierung und Investition, einschließlich Finanztechnik.

In jüngster Zeit hat sich Deep Learning aufgrund seiner weitreichenden Auswirkungen auf künstliche Intelligenz zu einer leistungsfähigen Maschinell-Lerntechnik entwickelt, obwohl Deep-Learning-Methoden derzeit nicht als allumfassende Lösung für die effektive Anwendung künstlicher Intelligenz betrachtet werden. Anns, die verschiedene Deep-Learning-Algorithmen verwenden, werden als Deep Neural Networks (dnns) kategorisiert, die in vielen wichtigen Bereichen angewendet wurden, wie automatische Spracherkennung, Bilderkennung, natürliche Sprachverarbeitung, Arzneimittelermittlung

und Toxikologie, Kundenbeziehungsmanagement, Empfehlungssysteme und Bioinformatik, wo sie oft gezeigt haben, dass sie verbesserte Ergebnisse für verschiedene Aufgaben liefern.

Darüber hinaus ist es für neuronale Netzwerke mit unterschiedlichen Topologien von entscheidender Bedeutung, mit einer bewussten Auswahl von Eingangsvariablen genaue Ergebnisse zu erzielen (Lam, 2004; Hussain et al., 2007). Die einflussreichsten und repräsentativsten Eingaben können unter anderem mit ausgereiften Dimensionsreduktionstechnologien wie derhauptkomponentenanalyse (PCA) und deren Varianten Fuzzy robust Principal Component Analysis (FRPCA) und Kernel-based Principal Component Analysis (KPCA) ausgewählt werden. PCA ist eine klassische und bekannte statistische lineare Methode, um die einflussreichsten Features aus einem hochdimensionalen Datenraum zu extrahieren. Van der Maaten et al. (2009) vergleichen PCA mit 12 front-ranked nonlinear dimensionality reduction techniques, wie multidimensionale Skalierung, Isomap, maximale Varianzentfaltung, KPCA, Diffusionskarten, mehrschichtige Autoencoder, lokal lineare Einbettung, Laplacian Eigenmaps, Hessian LLE, Local Tangent Space Analysis, locally linear coordination, and manifold charting, by applying each on self-created and natural tasks. Die Ergebnisse zeigen, dass, obwohl nichtlineare Techniken bei ausgewählten künstlichen Daten gut funktionieren, keine von ihnen die traditionelle PCA mit realen Daten übertrifft. Darüber hinaus erklären Sorzano, Vargas und Pascual-Montano (2014), dass PCA und seine Versionen, wie die Standard-PCA, robuste PCA, spärliche PCA und KPCA, nach wie vor für ihre Einfachheit und Intuition bevorzugt werden.

Nur wenige Studien haben sich auf die Vorhersage der täglichen Börsenrenditen mit hybriden Maschinell-Lernalgorithmen konzentriert. Zhong & Enke (2017a) präsentieren eine Studie zur Dimensionsreduktion mit einer Anwendung zur Vorhersage der täglichen Rücklaufrichtung des SPDR S&P 500 ETF (Tickersymbol: SPY) mit ANN-Klassifikatoren. Sie vergleichen verschiedene ANN-Modelle und stellen fest, dass unter den PCA- und ihren beiden beliebten Varianten FRPCA und KPCA PCA-basierte ANN-Klassifikatoren nachweislich der beste Prädiktor der ETF-Tagesrendite gegenüber verschiedenen Datensätzen sind, die mit PCA transformiert wurden (Zhong & Enke, 2017a). Darüber hinaus führt Zhong & Enke (2017b) ein umfassendes Data Mining-Verfahren durch, das sowohl Cluster- als auch Klassifizierungs-Mining einschließen, um die tägliche Renditerichtung des ETF vorherzusagen. Sie zeigen, dass PCA-basierte ANN-Klassifikatoren zu einer deutlich höheren Genauigkeit führen als drei verschiedene PCA-basierte logistische Regressionsmodelle, einschließlich derjenigen, die erfolgreich Fuzzy-C-Means-Clustering verwendet haben. Chong, Han, & Park (2017) untersucht kürzlich die Vor- und Nachteile der Verwendung von Deep Learning-Algorithmen für Aktienanalysen und -vorhersagen, aber ihre Studie konzentriert sich auf intraday-Aktienrendite-Prognosen.

In dieser Studie wird die tägliche Renditerichtung der SPDR S&P 500 ETF anhand eines bewusst entwickelten Klassifizierungsminingverfahrens auf Basis hybrider Maschinelearning-Algorithmen prognostiziert. Dieser Prozess beginnt mit der Vorverarbeitung der Rohdaten, um mit fehlenden Werten, Ausreißern und nicht übereinstimmenden Stichproben umzugehen. Die anns und dnns, die jeweils als Klassifikatoren fungieren, werden dann sowohl mit dem gesamten nicht transformierten Dataset als auch mit den PCA-dargestellten Datasets verwendet, um die Richtung zukünftiger täglicher Marktrenditen vorherzusagen. Der Rest dieses Papiers behandelt die Details der Studie und ist wie folgt organisiert. Als nächstes werden die Datenbeschreibung und die Vorverarbeitung eingeführt, einschließlich der Transformation des gesamten Datensatzes über PCA. Anschließend werden die Architekturen, Netzwerktopologie n- und Lernalgorithmen der neu entwickelten dnns sowie die zuvor erfolgreichen Benchmark-anns diskutiert, die beide für die Rückrichtungsklassifizierung verwendet werden. Anschließend werden die Prognoseprozeduren von drei verschiedenen Datensätzen mit den DNN-Klassifikatoren beschrieben, zusammen mit den Klassifizierungsergebnissen und dem Muster der Klassifizierungsgenauigkeit, die für die Anzahl der ausgeblendeten Layer relevant ist. Ein Standard-Benchmark wird auch mit den PCA-basierten ANN-Klassifikatoren-Ergebnissen verglichen. Die Simulationsergebnisse aus Handelsstrategien, die auf den DNN-Klassifikatoren über die drei Datensätze basieren, werden miteinander verglichen, und die Ergebnisse der ANN-basierten Handelsstrategien im Vergleich zu zwei Benchmarks werden dann diskutiert. Abschließend werden abschließende Bemerkungen und vorschläge für künftige Arbeiten vorgelegt.

SCHLUSSFOLGERUNG

Zur Vorhersage der täglichen Rücklaufrichtung des SPDR S&P 500 ETF (Tickersymbol: SPY) wurde ein umfassendes Big-Data-Analyseverfahren mit hybriden Machine Learning-Algorithmen entwickelt. Im Idealfall versuchen die Forscher, den einfachsten Satz von Algorithmen auf die geringste Datenmenge anzuwenden, wobei sowohl die genauesten Prognoseergebnisse als auch die höchsten risikobereinigten Gewinne gewünscht werden. Wir haben auch diesen Standard für diese Forschung berücksichtigt.

Der Analyseprozess beginnt mit der Datenreinigung und Vorverarbeitung und schließt mit einer Analyse der Prognose- und Simulationsergebnisse. Der Vergleich der Klassifizierungs- und Simulationsergebnisse erfolgt mit statistischen Hypothesentests, die zeigen, dass die Genauigkeit der DNN-basierten Klassifizierung im Durchschnitt signifikant höher ist als die PCA-dargestellten Daten über den gesamten nicht transformierten Datensatz. Genauer gesagt erreicht die DNN-basierte Klassifizierung für den PCA-dargestellten Datensatz mit pcs = 31 die höchste Genauigkeit. Es wird auch beobachtet, dass mit der Erhöhung der Anzahl der DNN-Verborgenenschichten ein Muster hinsichtlich der Klassifizierungsgenauigkeit (im Vergleich zum ANN-Klassifier) entsteht, wobei das Problem der Überanpassung unter Kontrolle bleibt. Darüber hinaus schneiden die Handelsstrategien mit den DNN-Klassifikatoren in den meisten Fällen bei drei Datensätzen mit unterschiedlichen Darstellungen besser ab als die, die die ANN-Klassifikatoren verwenden. Obwohl es im Allgemeinen keinen signifikanten Unterschied zwischen den Handelsstrategien aus dem DNN-Klassifizierungsprozess über den gesamten untransformierten Datensatz und zwei PCA-repräsentierte Datensätze gibt, schneiden die handelsbasierten Daten etwas besser ab.

In früheren Studien (Zhong & Enke, 2017a, 2017b) haben die PCA-ANN-Klassifikatoren eine höhere Vorhersagegenauigkeit für die tägliche Rücklaufrichtung der SPY ETF für den nächsten Tag als die FRPCA-ANN-Klassifikatoren, KPCA-ANN-Klassifikatoren und logistischen Regressionsklassifikatoren, mit oder ohne PCA/FRPCA/KPCA beteiligt. Auch die Handelsstrategien, die auf den PCA-ANN-Klassifikatoren basieren, schneiden besser ab als die anderen Strategien, die auf den anderen Klassifikatoren basieren. Darüber hinaus schneiden bei der Verwendung von PCA alle klassifizierungsmodellbasierten Handelsstrategien besser ab als die Benchmark-Einmonats-T-Bill-Strategie; die Handelsstrategien aus dem ANN-Klassifizierungsminingverfahren schneiden besser ab als die Benchmark-Buy-and-Hold-Strategie. In Kombination mit den neuen Ergebnissen, wie in den Tabellen 2, 3, 4 und 6, 7 8 dargestellt, kann daher der Schluss gezogen werden, dass unter den in dieser Studienreihe berücksichtigten Techniken des maschinellen Lernens die PCA-DNN-Klassifikatoren mit der richtigen Anzahl verdeckter Schichten die höchste Klassifizierungsgenauigkeit erreichen und zu der besten Handelsstrategieleistung führen können.

Mit zusätzlichen versteckten Schichten und komplizierteren Lernalgorithmen werden dnns als wichtige und fortschrittliche Technologie in den Bereichen Computerintelligenz und künstliche Intelligenz anerkannt. Dnns werden jedoch immer noch als Black Box mit weniger klaren theoretischen Bestätigungen der Lernalgorithmen betrachtet, die in gängigen tiefen Architekturen verwendet werden, wie z. B. Der stochastischen Gradientenabstiegsmethodik. Diese DNN-Lernalgorithmen erhöhen tatsächlich die Rechenzeit, da eine große Anzahl von versteckten Schichten und Neuronen enthalten sind. Dieser Forschungsbereich muss in Zukunft mehr Aufmerksamkeit und Anstrengungen erhalten.

TRANSLATED VERSION: PORTUGUESE

Below is a rough translation of the insights presented above. This was done to give a general understanding of the ideas presented in the paper. Please excuse any grammatical mistakes and do not hold the original authors responsible for these mistakes.

VERSÃO TRADUZIDA: PORTUGUÊS

Aqui está uma tradução aproximada das ideias acima apresentadas. Isto foi feito para dar uma compreensão geral das ideias apresentadas no documento. Por favor, desculpe todos os erros gramaticais e não responsacule os autores originais responsáveis por estes erros.

INTRODUÇÃO

Técnicas analíticas de big data desenvolvidas com algoritmos de machine learning estão ganhando mais atenção em vários campos de aplicação, incluindo investimento no mercado de ações. Isso ocorre principalmente porque os algoritmos de aprendizagem de máquina não exigem suposições sobre os dados e muitas vezes alcançam maior precisão do que os modelos econométricos e estatísticos; por exemplo, redes neurais artificiais (anns), sistemas difusos e algoritmos genéticos são impulsionados por dados multivariados sem suposições necessárias. Muitas dessas metodologias têm sido aplicadas para prever e analisar variáveis financeiras, por exemplo, ver Vellido, Lisboa e Meehan (1999); Kim & Han (2000); Cao & Tay (2001); Thawornwong, Dagli e Enke (2001); Bogullu, Enke e Dagli (2002); Hansen & Nelson (2002); Wang (2002); Chen, Leung, & Daouk (2003); Zhang (2003); Chun & Kim (2004); Shen & Loh (2004); Thawornwong & Enke (2004); Armano, Marchesi, & Murru (2005); Enke & Thawornwong (2005); Ture & Kurt (2006); Amornwattana et al. (2007); Enke & Mehdiyev (2013); Zhong & Enke (2017a, 2017b); Huang & Kou (2014); Huang, Kou e Peng (2017); e Nayak & Misra (2018). Uma revisão abrangente desses estudos foi conduzida por Atsalakis & Valavanis (2009) e Vanstone & Finnie (2009). Com características não lineares, baseadas em dados e fáceis de generalizar, a análise multivariada com anns tornou-se uma ferramenta dominante e popular de análise em finanças e economia. Refenes, Burgess, & Bentz (1997) e Zhang, Patuwo, & Hu (1998) revisam o uso do uso de anns como método de previsão em diferentes áreas de finanças e investimentos, incluindo engenharia financeira.

Recentemente, o deep learning surgiu como uma poderosa técnica de aprendizado de máquina devido às suas implicações de longo alcance para a inteligência artificial, embora os métodos de deep learning não sejam atualmente considerados como uma solução abrangente para a aplicação eficaz da inteligência artificial. As anns que utilizam diferentes algoritmos de aprendizagem profunda são categorizadas como redes neurais profundas (dnns), que têm sido aplicadas a muitos campos importantes, como reconhecimento automático de fala, reconhecimento de imagem, processamento de linguagem natural, descoberta e toxicologia de medicamentos, gerenciamento de relacionamento com o cliente, sistemas de recomendação e bioinformática, onde muitas vezes têm sido mostrados para produzir resultados aprimorados para diferentes tarefas.

Além disso, é fundamental que redes neurais com diferentes topologias alcancem resultados precisos com uma seleção deliberada de variáveis de entrada (Lam, 2004; Hussain et al., 2007). Os insumos mais influentes e representativos podem ser escolhidos utilizando tecnologias maduras de redução de dimensionalidade, como a análise de componentes principais (PCA), e suas variantes fuzzy robust principal component analysis (FRPCA) e análise de componentes principais baseados em kernel (KPCA), entre outros. PCA é um método linear estatístico clássico e bem conhecido para extrair as características mais influentes de um espaço de dados de alta dimensão. Van der Maaten et al. (2009) comparam o PCA com 12 técnicas de redução de dimensionalidade não linear, como dimensionamento multidimensional, Isomap, desdobramento de variância máxima, KPCA, mapas de difusão, autoencoders multicamadas, incorporação localmente linear, eigenmaps laplacianos, LLE hessian, análise de espaço tangente local, coordenação local linear e gráficos múltiplos, aplicando cada um em tarefas auto-criadas e naturais. Os resultados mostram que, embora as técnicas não lineares tenham um bom desempenho em dados artificiais selecionados, nenhuma delas supera o PCA tradicional usando dados do mundo real. Além disso, Sorzano, Vargas e Pascual-Montano (2014) afirmam que entre as técnicas de redução de dimensionalidade disponíveis, o PCA e suas versões, como o PCA padrão, PCA robusto, PCA esparso e KPCA, ainda são preferidos por sua simplicidade e intuitiva.

Poucos estudos se concentraram em prever retornos diários do mercado de ações usando algoritmos híbridos de aprendizado de máquina. Zhong & Enke (2017a) apresentam um estudo de redução de dimensionalidade com um aplicativo para prever a direção de retorno diário do ETF SPDR S&P 500 (símbolo do ticker: SPY) usando classificadores ANN. Eles comparam vários modelos ANN e descobrem que entre o PCA e suas duas variantes populares, FRPCA e KPCA, classificadores ANN baseados em PCA são mostrados como o melhor preditor da direção de retorno diário do ETF sobre vários conjuntos de dados transformados usando PCA (Zhong & Enke, 2017a). Além disso, a Zhong & Enke (2017b) realiza um procedimento abrangente de mineração de dados, incluindo a mineração de cluster e classificação, para prever a direção de retorno diário do ETF. Eles mostram que os classificadores ANN baseados em PCA levam a uma precisão significativamente maior do que três modelos diferentes de regressão logística baseada em PCA, incluindo aqueles que usaram com sucesso clustering c-means fuzzy. Chong, Han e Park (2017) examinam recentemente as vantagens e desvantagens do uso de algoritmos de aprendizagem profunda para análise e previsão de estoque, mas seu estudo se concentra na previsão intradiária de retorno das ações.

Neste estudo, a direção de retorno diário do ETF SPDR S&P 500 é prevista usando um procedimento de mineração de classificação deliberadamente projetado com base em algoritmos híbridos de aprendizagem de máquina. Esse processo começa pré-processando os dados brutos para lidar com valores perdidos, outliers e amostras incompatíveis. As anns e dnns, cada uma atuando como classificadores, são então usadas com todo o conjunto de dados não transformado e os conjuntos de dados representados pelo PCA para prever a direção dos retornos futuros do mercado diário. O restante deste artigo discute os detalhes do estudo e é organizado da seguinte forma. A descrição dos dados e o pré-processamento são introduzidos em seguida, incluindo a transformação de todo o conjunto de dados via PCA. As arquiteturas, topologia de rede e algoritmos de aprendizagem dos dnns recém-desenvolvidos, juntamente com as anns de benchmark de sucesso anteriormente bem sucedidas, ambas usadas para classificação de direção de retorno, são então discutidas. O procedimento de previsão de três conjuntos de dados diferentes com os classificadores DNN são então descritos, juntamente com os resultados de classificação e o padrão da precisão de classificação relevante para o número de camadas ocultas. Um benchmark padrão também é comparado com os resultados de classificadores ANN baseados em PCA. Os resultados da simulação a partir de estratégias de negociação baseadas nos classificadores DNN ao longo dos três conjuntos de dados são comparados entre si, e os resultados das estratégias de negociação baseadas em ANN em comparação com dois benchmarks são então discutidos. Finalmente, são fornecidas observações finais e trabalhos futuros propostos.

CONCLUSÃO

Um procedimento abrangente de análise de big data usando algoritmos híbridos de aprendizado de máquina foi desenvolvido para prever a direção de retorno diário do ETF SPDR S&P 500 (símbolo de ticker: SPY). Idealmente, os pesquisadores procuram aplicar o conjunto mais simples de algoritmos à menor quantidade de dados, com os resultados de previsão mais precisos e os maiores lucros ajustados ao risco sendo desejados. Também consideramos este padrão para esta pesquisa.

O processo analítico começa com limpeza e pré-processamento de dados e conclui com uma análise dos resultados de previsão e simulação. A comparação dos resultados de classificação e simulação é feita com testes estatísticos de hipóteses, mostrando que, em média, a precisão da classificação baseada em DNN é significativamente maior do que os dados representados pelo PCA sobre todo o conjunto de dados não formulados. Mais especificamente, a classificação baseada em DNN para o conjunto de dados representado pelo PCA com pcs = 31 alcança a maior precisão. Observa-se também que, à medida que o número de camadas ocultas de DNN aumenta, surge um padrão em relação à precisão de classificação (em comparação com o classificador ANN), com o problema de sobrejustificação permanecendo sob controle. Além disso, ao longo de três conjuntos de dados com diferentes representações, as estratégias de negociação que utilizam os classificadores DNN têm um desempenho melhor do que as que usam os classificadores ANN na maioria dos casos. Embora, em geral, não haja diferença significativa entre as estratégias de negociação do processo de classificação DNN ao longo de todo o conjunto de dados não transformados e dois conjuntos

de dados representados pelo PCA, as estratégias de negociação baseadas nos dados representados pelo PCA têm um desempenho ligeiramente melhor.

Em estudos anteriores (Zhong & Enke, 2017a, 2017b), os classificadores PCA-ANN são mostrados para dar uma maior precisão de previsão para a direção de retorno diário do ETF spy para o dia seguinte do que os classificadores FRPCA-ANN, classificadores KPCA-ANN e classificadores de regressão logística, com ou sem PCA/FRPCA/KPCA envolvidos. Além disso, as estratégias de negociação baseadas nos classificadores PCA-ANN têm um desempenho melhor do que as outras estratégias baseadas nos outros classificadores. Além disso, ao usar o PCA, todas as estratégias de negociação baseadas em modelos de classificação têm um desempenho melhor do que a estratégia de benchmark de um mês de conta T; as estratégias de negociação do procedimento de mineração de classificação ANN têm um desempenho melhor do que a estratégia de compra e resarciência de benchmark. Assim, quando combinado com os novos resultados ilustrados nas Tabelas 2, 3, 4 e 6, 7 8 pode-se concluir que entre as técnicas de aprendizado de máquina consideradas nesta série de estudo, os classificadores PCA-DNN com o número adequado de camadas ocultas podem alcançar a maior precisão de classificação e resultar no melhor desempenho da estratégia de negociação.

Com camadas ocultas adicionais e algoritmos de aprendizagem mais complicados, os dnns são reconhecidos como uma tecnologia importante e avançada nos campos da inteligência computacional e inteligência artificial. No entanto, os dnns ainda são considerados como uma caixa preta com confirmações teóricas menos claras dos algoritmos de aprendizagem que são usados em arquiteturas profundas comuns, como a metodologia de descida de gradiente estocástico. Esses algoritmos de aprendizagem DNN realmente aumentam o tempo de computação à medida que um grande número de camadas ocultas e neurônios estão incluídos. Essa área de pesquisa precisa receber mais atenção e esforço no futuro.