Journal of the Korean Institute of Industrial Engineers
Vol. 37, No. 1, pp. 64-73, March 2011.

기계학습기법에 기반한 국제 유가 예측 모델

박강희[1] · Tianya Hou[1] · 신현정[2†]

[1]아주대학교 산업공학과 / [2]아주대학교 산업정보시스템공학부

Oil Price Forecasting Based on Machine Learning Techniques

Kanghee Park[1] · Tianya Hou[1] · Hyunjung Shin[2]

[1]Dept. of Industrial Engineering Ajou University
[2]Dept. of Industrial and Information System Engineering, Ajou University

Oil price prediction is an important issue for the regulators of the government and the related industries. When employing the time series techniques for prediction, however, it becomes difficult and challenging since the behavior of the series of oil prices is dominated by quantitatively unexplained irregular external factors, e.g., supply- or demand-side shocks, political conflicts specific to events in the Middle East, and direct or indirect influences from other global economical indices, etc. Identifying and quantifying the relationship between oil price and those external factors may provide more relevant prediction than attempting to unclose the underlying structure of the series itself. Technically, this implies the prediction is to be based on the vectoral data on the degrees of the relationship rather than the series data. This paper proposes a novel method for time series prediction of using Semi-Supervised Learning that was originally designed only for the vector types of data. First, several time series of oil prices and other economical indices are transformed into the multiple dimensional vectors by the various types of technical indicators and the diverse combination of the indicator-specific hyper-parameters. Then, to avoid the curse of dimensionality and redundancy among the dimensions, the well-known feature extraction techniques, PCA and NLPCA, are employed. With the extracted features, a timepoint-specific similarity matrix of oil prices and other economical indices is built and finally, Semi-Supervised Learning generates one-timepoint-ahead prediction. The series of crude oil prices of West Texas Intermediate (WTI) was used to verify the proposed method, and the experiments showed promising results : 0.86 of the average AUC.

Keywords: Oil Price Prediction, Time Series, Technical Indicators, Feature Extraction(PCA/NLPCA), Semi-Supervised Learning(SSL)

1. 서 론

원유의 가격은 정부 정책 및 관련 산업 등 국가 경제 전반에 중대한 영향을 미친다. 구체적으로 에너지, 운송업, 주거 및 섬유 산업까지 인간이 삶을 영위하는데 없어서는 안될 생활 필수품이다. 이런 이유로 세계 각국은 유가에 관심을 보이고 국제 원유의 거래 규모는 세계 무역의 10%를 차지한다(Verleger, 1993). 유가의 등락은 모든 나라들의 경제활동에 민감한 영향을 미치는데 유가가 올라가면 OECD와 개발 도상국 같은 석유수입국은 경제 전체의 물가가 상승해 실질적인 경기 침체로 이어지고, 반면에 1998년 같이 유가가 하락하게 되면 중동지역의 주요 석유 수출국은 심각한 경제 적자를 초래하게 된다(Birol, 2004; Abosedra and Baghestan, 2004). 따라서 유가 등락을 예측하는 것은 과거부터 계속되어온 전세계적인 이슈 중 하나이다(Stevens, 1995).

그러나 유가를 예측하는 것은 현실적으로 상당히 어려운 문

본 연구는 2008년 에너지 경제 연구원의 데이터를 기반으로 수행되었으며, 아주대학교 Post BK 21 연구비지원 및 한국연구재단(2010-0007804)의 연구비 지원으로 수행되었음을 사사하며, 이에 감사의 뜻을 표한다.
† 연락저자 : 신현정 교수, 443-749 경기 수원시 영통구 원천동 아주대학교 산업정보시스템공학부, Tel : 031-219-2417, Fax : 031-219-1610, E-mail : shin@ajou.ac.kr
2010년 12월 6일 접수; 2011년 1월 20일 수정본 접수; 2011년 2월 15일 게재 확정.

제이다. 유가 변동에 영향을 미치는 비정량적, 비정규적, 설명하기 어려운 외부 요인들(수요-공급의 갑작스러운 파동, 중동 지역의 정치적인 분쟁, 다른 국제 경제지표의 직-간접적인 영향 등)과의 상호 영향 관계 등을 모두 고려해야 하기 때문이다(Lynch, 2003; Basher and Sadorsky, 2006; Amano and Norden, 1998; Svensson, 2005; He *et al*., 2009).

그래서 유가 예측 방법은 과거로부터 다양한 통계적 또는 경제적 분석방법 등의 형태로 지속적으로 발전되어 왔다. 유가 데이터는 시간의 흐름, 계절순환에 따른 시계열 데이터 특성을 가지므로, wavelet-based 모델, semi-parametric 모델, 금융 시계열 모델 등 시계열 분석기법들이 분석의 주된 도구로 사용되어 왔다(Keong, 2009). 예를 들어 Yousefi *et al*.(2005)의 논문에서는 wavelet-based 알고리즘을 사용하여 유가 가격을 예측하였고 Morana(2001)은 GARCH 모델을 기본으로 semi-parametric 통계 기법을 사용하여 유가 단기 예측을 제시하였다. Cortazar and Schwartz(2002)는 유가 예측에 금융분석 모델을 접목시켰다. 이 방법들은 과거의 유가 등락을 기초로 한 시계열 데이터로 분석을 한 것이기 때문에 미래의 유가가 과거와 유사하게 움직일 것이라는 가정을 토대로 한다. 그러나 위 연구들에서는, 유가는 그 특성상 다른 경제 지표들과의 상호작용이 유가의 등락에 영향을 미친다는 점을 분석에 충분히 반영하지 못했다는 아쉬움이 있다(Basher and Sadorsky, 2006). 이를 반영한 연구들에서는 autoregressive integrated moving average (ARIMA) models, error correction models(ECMs), vector auto-regression(VAR) models 등을 사용하였다. Akarca and Andrianacos(1997)은 ARIMA model을 이용해 유가예측을 제시하였고 Lanza *et al*.(2005)은 석유 생산 가격을 결합한 error correction models(ECMs)을 이용해 유가를 조사하였다. Mirmirani and Li (2004)은 vector auto-regression(VAR)를 이용하여 미국 내의 유가예측을 제안하였다.

위에서 제시한 기존연구들의 대부분은 여전히 시계열 분석에 기반하여 단순히 모델링에 이용하거나 혹은 관련 있는 경제 지표를 모델에 결합시키는 방법으로 예측을 수행하였다. 그런데 유가 예측에는 유가 변동에 영향을 미치는 비정량적, 비정규적, 설명하기 어려운 외부 요인들과의 상호 영향 관계가 보다 명확히 고려되어야 한다. 그러나 앞선 연구사례에서 사용되어왔던 시계열 분석방법만으로는 이러한 상호 관계 및 인과관계를 명시적으로 모델에 포함하고 정형화하는 것에 방법론적으로 한계가 있다. 따라서 시계열 데이터를 일반 벡터 타입의 데이터로 변환시킨 후 이를 외부 요인과의 인과관계를 통하여 예측하는 방법이 필요하다. 벡터 타입의 분석을 위해 최근 artificial neural network(ANN)와 support vector machine (SVM)을 이용한 방법들이 제안되고 있다. 예를 들어 Yu *et al*. (2008)의 논문에서는 EMD-based neural network ensemble learning paradigm을 이용하여 유가예측을 시도하였고, Xie *et al*. (2006)의 논문에서는 support vector machine(SVM)을 이용하여 유가를 예측하였다. 그러나 ANN과 SVM의 방법에도 결점은 존재한다. ANN은 매번 같은 시점의 예측값이 계속 달라지는 경향이 있고, SVM의 경우는 커널함수를 이용하여 시계열 데이터를 벡터 타입으로 변환시키는 방법이 있으나 아직까지 그 연구의 깊이가 미미하다. 또한 ANN과 SVM은 그 특성상 유가와 수요, 유가와 공급, 유가와 기타 경제 지표간의 직접적인 상호관계를 고려 할 수는 있으나 수요와 공급 또는 수요와 기타 경제지표 등 입력 변수들끼리의 상호작용을 분석하기는 어렵다.

이러한 결점을 극복하기 위해 본 논문에서는 최근 기계학습 분야에서 주목 받고 있는 semi-supervised learning(SSL)을 유가 예측에 적용하는 방법론을 제안하고자 한다(Shin *et al*., 2007; Shin *et al*., 2010). SSL 기법을 시계열 분석에 적용 한 사례로는 Cheng and Tan(2008)의 날씨예측에 관한 연구가 있으나 이 연구는 시계열 데이터를 그대로 사용한 분석이었다(Cheng and Tan, 2008). 본 논문에서는 시계열 데이터를 벡터 형태로 전환한 후 SSL에 접목시켜 유가를 예측하는 모델을 제안한다. 제안한 모델은 1992년 1월부터 2008년 6월의 서부 텍사스 중질유(West Texas Intermediate, WTI) 유가 등락 분석에 적용, 검증되었다.

2. 방법론

유가 예측 모델은 수요부분, 공급부문, 기타 요인들을 평가하고 그에 따라 유가의 등락 및 석유시장의 종합적인 현황을 고려해야 한다. 이에 본 논문에서 제안하는 모델은 다음과 같은 절차를 따른다. 첫째로, 석유시장에 관련된 여러 자료들을 수집하여 정리하고 정확성을 확인하는 전처리 단계를 수행한다. 모든 고려되는 변수들과 유가(WTI)와의 관계에 대해 통계분석을 통하여 관련성을 정량화하고 이들 중 관련성이 높은 변수들을 선별한다. 둘째, 선택된 입력변수들의 시계열 데이터를 벡터 타입의 데이터로 변환시키기 위하여 기술적 지표(technical indicators, TI)를 이용한다. 셋째, 각 기술적 지표들은 사용자가 설정해 주어야만 하는 파라미터 값이 있는데, 이 값을 어떻게 설정하느냐에 따라 다수 개의 파생변수들이 생성된다. 즉, 하나의 지표에서 다수 개의 파생변수들이 발생한다. 변수의 차원이 증가하면 데이터 잡음이 증가하고 예측모델의 과적합(overfitting)이 발생할 수 있다. 이를 줄여주기 위해 선형 또는 비선형 변수변환을 통하여 한 개 또는 두 개의 대표적 특성변수를 추출한다. 본 연구에서는 선형 및 비선형 변수추출 방법으로 잘 알려진 주성분분석(Principal Component Analysis, PCA) 및 비선형주성분분석(Nonlinear Principal Component Analysis, NLPCA)을 각각 사용한다. 넷째, 벡터 타입으로 변환된 변수들을 Semi-Supervised Learning(SSL)을 이용하여 WTI 유가 예측모델을 만든다.

다음 각 절에서는 semi-supervised learning, 기술적 지표(Technical Indicators), 변수추출방법(주성분분석, 비선형주성분분석)에 대하여 간략히 소개한다.

2.1 Semi-Supervised Learning(SSL)

Semi-supervised learning은 레이블이 알려진 데이터와 레이블이 알려지지 않은 데이터를 함께 고려하여 학습을 수행하는 방법이다. 잘 알려진 알고리즘 중의 하나인 Graph-based Semi-Supervised Learning(GSSL)에서는 그래프를 이용하여 유사도가 높은 데이터들을 클러스터링 하고 이를 토대로 레이블이 없는 데이터의 레이블을 예측한다(Shin *et al*., 2007). <그림 1>는 GSSL의 기본 모형을 보여준다.

GSSL 방법은 전체 데이터 n(n = l+u)개 중에서 l개의 레이블이 있는 데이터 $\{(X_1, Y_1), \cdots, (X_l, Y_l)\}$와 u개의 레이블이 정해지지 않은 데이터를 사용한다. 데이터의 레이블은 레이블이 있을 경우 -1 또는 1로 설정하며($Y_l \in \{-1, 1\}$, l = 1, ⋯, l) 그렇지 않은 경우 0으로 설정한다($Y_u \in \{0\}$, u = l+1, ⋯, l+u).

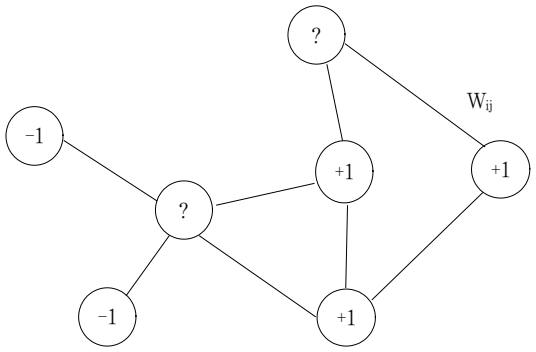

Figure 1. Graph-based semi-supervised learning(SSL)

데이터들은 그래프의 노드들로 표현되고 유사도가 높다고 판단되는 노드들 사이에는 엣지가 형성된다. i노드와 j노드 사이의 엣지의 연결강도는 유사도가 증가할수록 커지는데, 다음의 식 (1)에 의해 결정된다.

$$w_{ij} = \begin{cases} \exp\left(-\dfrac{(x_i - x_j)^T (x_i - x_j)}{\sigma^2}\right) & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

유사행렬 W를 만드는 방법으로는 주로 k-nearest neighbor (KNN) 또는 일정 반경 내에서의 유클리드 거리를 활용한 방법이 주로 사용된다($\| x_i - x_j \|^2 < r$, euclidean distance within a certain radius). SSL 알고리즘은 식 (2)의 이차목적함수를 풂으로써 레이블이 없는 노드들에 대한 예측값 f를 출력한다.

$$\mathrm{Min}_f\, (f-y)^T (f-y) + \mu f^T L f \quad (2)$$

여기서 y는 목표값으로서 $y = (y_1, \cdots, y_l, 0, \cdots, 0)^T$로 설정되며 출력값(예측값)은 $f = (f_1, \cdots, f_l, f_{l+1}, \cdots, f_{n=l+u})^T$로 표현된다. **L**은 graph laplacian matrix이며 **D** = diag(d_i), **L** = **D**-**W**로 정의된다. 알고리즘은 두 가지 학습조건인, (a) 출력값 f 는 레이블이 된 노드에서는 노드의 목표 값 y와 비슷해야 하고, (b) 연결된 두 노드 I와 j의 출력값은 크게 달라지면 안 된다는 제약을 만족시키는 해를 산출한다. 식 (2)의 파라미터 μ는 이러한 두 가지 조건이 학습에 미치는 영향을 조절하는 학습 파라미터로서 사용자에 의해 정의된다. 식 (2)로부터 다음의 식 (3)을 유도하여 출력값 f 를 계산할 수 있다.

$$f = (I + \mu L)^{-1} y \quad (3)$$

I는 단위행렬을 의미한다.

SSL은 입력값으로 벡터 타입의 변수를 사용하는데 유가 데이터는 시계열 형태의 데이터로 존재하기 때문에 직접적으로 유가 데이터를 SSL에 적용하기 어렵다. 따라서 다음 장에서 시계열 데이터를 벡터 타입으로 변환시키는 방법을 제시한다.

2.2 기술적 지표 (Technical Indicators, TI)

기술적 분석은 금융분석에서 자주 사용하는 방법인데 기술적 지표화의 장점은 시계열 데이터가 고유하게 갖고 있는 잡음(oscillatory noise)을 제거하고 그 밑에 존재하는 큰 구조, 즉 추세 및 구조적 변동요인들을 이끌어 내는 역할을 한다. 유가와 경제지표들은 변수 특성상 시계열 데이터로 존재한다. 이 시계열 데이터는 시간에 따른 연관관계를 형성하고 식 (4)와 같이 표현된다.

$$X_t = \{x_1, x_2, \cdots, x_i, \cdots, x_t\} \quad (4)$$

여기서 x_i는 t를 전체 시간으로 할 때 i번째 시점의 데이터를 의미한다. X_t가 시계열 데이터인 특성상 SSL에 직접 적용하기에는 몇 가지 문제점이 존재한다. <그림 2>에서 보면 그래프의 전체 노드들은 식 (4)와 같이 각 노드마다의 고유 시계열 데이터를 갖고 있다.

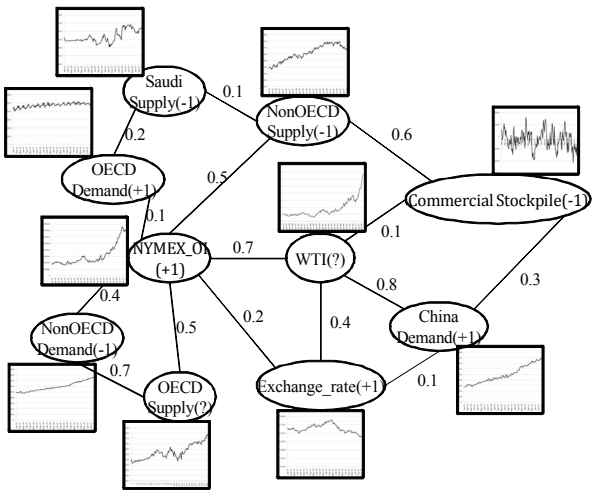

Figure 2. Graph SSL representation for time series prediction

예를 들어 <그림 2>에서 WTI 노드도 시계열 데이터 X_t^{WTI}가 있고 Saudi 노드 또한 고유의 시계열 데이터 X_t^{Saudi}가 있다. 문제는, 이 두 시계열 데이터들은 시계열 데이터 자체 고유의 잡음을 포함하고 있고, 둘 사이의 유사도를 도출하는데 잡음을 포함하는 문제점이 있다. 따라서 각 시계열 데이터들을 구조상의 요인들 즉 추세, 변동 요인들로 변환하여 벡터화하고 이로부터 두 노드, WTI와 Saudi 간의 유사도를 측정한다. 다음 <표 1>은 본 연구에서 시계열 데이터를 7가지 벡터타입의 데이터로 변환시키는데 사용한 기술적 지표들을 정리한 것이다.

Table 1. The Definition of Technical Indicators(TIs)

	지표	설명
s_1	$MA_z(X_t)=\frac{1}{z}(x_t)+\frac{z-1}{z}MA_z(X_{t-1})$	z-이동평균값 (평활값)
s_2	$BLAS_z(X_t)=\frac{x_t-MA_z(X_t)}{MA_z(X_t)}$	현재값과 이동평균값 차이의 상대비율
s_3	$OSC_{j,z}(X_t)=\frac{MA_j(X_t)-MA_z(X_t)}{MA_z(X_t)}$	이동평균의 최고 최저점 비율
s_4	$ROC_z(X_t)=\frac{x_t-x_{t-z}}{x_t}$	현재 값과 z시점 이전 값과의 상대 차이
s_5	$K_n^z=\frac{C_n-Min_{i=n-z-1}^{n}(L_i)}{Max_{i=n-z-1}^{n}(H_i)-Min_{i=n-z-t}^{n}(L_i)}$	Cn : 구간내 최종가, Ln : 구간내 최저가, Hi : 구간내 최고가
s_6	$D_n^z=MA_3(K_n^z)$,	K의 3달 이동평균선
s_7	$RSI_t^z=\frac{\sum_{i=t-z-1,x_i>x_{i-1}}^{t}(\lvert x_i-x_{i-1}\rvert)}{\sum_{i=t-z-1}^{t}(\lvert x_i-x_{i-1}\rvert)}$,	상승 압력과 하락 압력간의 상대적인 강도

위 지표들을 이용하면 시계열 데이터의 시간 연관성을 보존하면서도 벡터 타입의 데이터로 변환할 수 있으므로 SSL의 적용이 용이하게 된다. <표 1>에서 보듯이 각 지표들에는 사용자가 결정해야 할 파라미터들이 존재한다. 그러나 적절한 파라미터 값을 결정하는 일은 어떠한 규칙이 존재하지 않으므로 대개는 시행착오(trial-and-error)를 거쳐 이루어진다. 또 다른 대안으로는 다양한 파라미터 값들을 모두 고려하는 방법이다. 그러나 이러한 경우에는, 파라미터 값의 개수에 따라 하나의 지표가 여러 개의 변수로 늘어나게 되고, 이는 입력변수의 차원을 증가시키게 하여 고차원 문제(curse of dimensionality)를 유발할 수 있으며 모델의 과적합을 일으킬 소지가 다분하다. 따라서 다음 장에서는 다양한 파라미터 값을 모두 사용하되 이로부터 파생된 불필요한 차원을 감소시키는 방법에 대하여 소개한다.

2.3 변수 추출 및 차원 감소

2.3.1 주성분분석 (Principal Component Analysis)

Principal component analysis(PCA)의 장점은 선형 변환을 이용하여 원 데이터의 차원을 감소시키는 역할을 한다. PCA는 데이터들을 한 개의 축으로 사상시켰을 때 분산이 가장 커지는 축을 첫 번째 좌표축으로, 두 번째 커지는 축을 두 번째 좌표축으로 놓는 방식으로, 새로운 좌표계를 형성하여 데이터를 선형 변환 시키는 방법이다. 차원이 m개인 입력데이터 $x_j \in R^m$ $(i=1,\cdots,n, \sum_{i=1}^{n} x_i = 1, m<n)$가 주어졌을 때, 새로운 좌표계로 사상시켜 s_i로 만드는 PCA 선형 변환식 (5)은 다음과 같다.

$$\underbrace{s_i}_{m\times 1} = \underbrace{U^T}_{m\times m}\underbrace{x_i}_{m\times 1},\ i=1,\cdots,n \tag{5}$$

여기서 U는 $m\times m$의 정방행렬이고 k번째 열 u_k는 공분산행렬(covariance matrix) $C=\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$의 k번째 고유벡터(eigenvector)이다. 행렬 U는 공분산행렬 C에 대한 고유값(eigenvalue) 문제를 다음의 식 (6)에서처럼 풂으로서 얻을 수 있다.

$$\lambda_k u_k = C u_k,\ k=1,\cdots,m \tag{6}$$

여기서 λ_k는 C의 고유값이고 u_k는 고유벡터이다. 새로운 좌표계에서 분산의 크기는 고유값으로 설명할 수 있다. 그러므로 고유값들을 크기에 따라 내림차순으로 정렬시킨 후, $\lambda_1 > \lambda_2 > \cdots > \lambda_p > \cdots > \lambda_m$, 첫 번째 p개의 고유값과 이와 대응되는 p개의 고유벡터 $\tilde{U}^T = \{u_1, u_2, \cdots, u_p\}$로 새로운 좌표계를 형성하면, 데이터의 중요한 특징을 유지하면서도 새로운 p차원의 직교공간(orthonomal space)으로 차원을 감소시키는 효과를 거둘 수 있다. 식 (5)로부터 m차원의 데이터 x_i를 p차원의 데이터 $\tilde{s}_i$로 직교변환시키는 공식을 다음과 같이 얻을 수 있다.

$$\underbrace{\tilde{s}_i}_{p\times 1} = \underbrace{\tilde{U}^T}_{p\times m}\underbrace{x_i}_{m\times 1},\ i=1,\cdots,n \tag{7}$$

PCA는 차원을 감소시키는 우수한 방법이나 이를 활용하려면 변수들이 모두 선형관계를 가져야 한다는 제한 조건이 있다. 따라서 데이터가 비선형관계를 갖는다면, 이러한 경우에는 PCA로 중요한 특성변수를 추출하는 것이 어려우므로 입력변수 차원감소의 효과를 얻는 것 역시 어렵다.

2.3.2 비선형주성분분석(Nonlinear Principal Component Analysis) : 자기연상신경망(Autoassociative Neural Network)

데이터의 차원을 감소시키는 다른 방법으로 autoassociative

neural network(AANN)의 방법이 있다. AANN은 전방향신경망(feed-forward NN)의 일종으로 데이터가 비선형관계일 때도 사용할 수 있는 변수변환 및 차원감소에 사용되는 알고리즘이다. AANN의 일반적인 구조는 <그림 3>와 같다. AANN은 입력층과 출력층, 그리고 은닉층으로 이루어져 있다. 입력노드들과 출력노드들의 개수는 각각 m개로 입력데이터 차원과 같다. 은닉층은 mapping function(F1)을 구현하는 mapping layer와 de-mapping function(F2)를 구현하는 demapping layer, 그리고 차원감소를 수행하는 bottleneck layer로 구성된다. Bottleneck layer의 노드 수(p)는 입력층과 출력층의 노드 숫자보다 적게 설정하는데($p < m$), 이는 m차원의 원 데이터가 p차원의 데이터로 변환되는 차원감소를 구현하기 위해서다.

Autoassociative mapping에서는 목표 데이터가 입력 데이터와 동일하게 설정된다. 이를 "identity mapping"라 하며 데이터가 입력층으로부터 출력층으로 나오기 전에 bottleneck layer를 통과하게 함으로써 bottleneck layer의 은닉노드들로 하여금 데이터의 주요한 특징들을 습득할 수 있도록 한다. 이를 수식으로 표현하면 다음과 같다. F를 학습된 autoassociative mapping이라 하고 주어진 입력데이터를 $\{x_1, x_2, \cdots, x_n\}$, AANN에 의하여 나온 출력값을 $\{\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n\}$이라고 하자. 그러면 다음의 식 (8)에서 처럼 학습오차 E를 최소화하는 과정에서 F를 얻을 수 있다.

$$E = \sum_{i=1}^{n} (x_i - \tilde{x}_i)^T (x_i - \tilde{x}_i) = \sum_{i=1}^{n} (x_i - F(x_i))^T (x_i - F(x_i)) \quad (8)$$

Mapping function F는 F1과 F2로 나눌 수 있으며, $F(\cdot) = F_2(F_1(\cdot))$로 정의된다. F1은 네트워크를 통해 입력층이 더 낮은 차원의 bottleneck layer로 변환되는 과정을 구현한 것이며 F2는 bottleneck layer로부터 출력층으로, 즉 원 데이터의 입력값을 복원하는 과정을 구현한다.

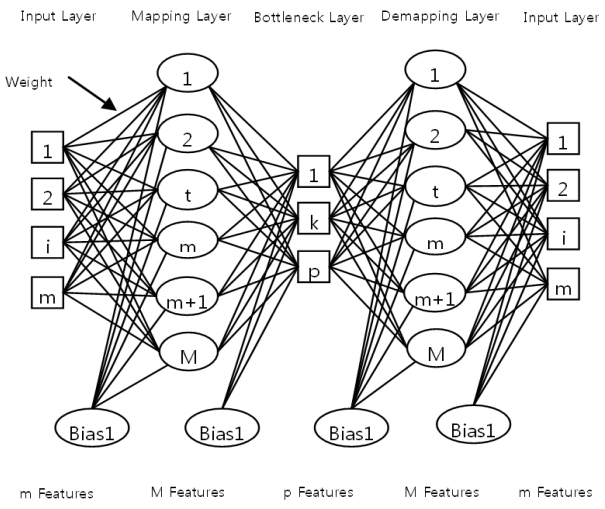

Figure 3. Architecture of AANN

요약하면 데이터는 더 낮은 차원으로 압축되었다가, 재조직화된다. 입력층에서 bottleneck layer로 mapping하는 것은 낮은 차원의 공간(m -> p)으로 "비선형사영"하는 것으로 여겨진다. 이러한 bottleneck layer의 노드들은 데이터의 중요한 정보들을 유지한 채 변수를 추출하는데 고려될 수 있다. 새로운 저차원 공간에서의 데이터 $\tilde{s}_i$는 식 (9)과 같이 계산된다.

$$\underset{p\times 1}{\tilde{s}_i} = F_1 \underset{m\times 1}{(x_i)} \quad (9)$$

AANN은 그 구조가 잘 설정된다면 데이터의 비선형관계를 잘 표현할 수 있는 변수를 추출할 수 있는 좋은 모델이다. 그러나 AANN은 bottleneck layer의 노드 개수가 처음부터 정해져 있는 방법은 아니므로 상황에 맞게 사용자가 결정하여야 한다는 어려움이 있다.

3. 실험

3.1 데이터

본 실험에서 사용한 데이터는 서부 텍사스 중질유(WTI) 가격에 대한 시계열 데이터로서, 이는 국내유가의 가격형성에 가장 큰 영향을 미친다고 알려진 국제유가이다. 데이터는 <그림 4>에서와 같이 1992년 1월부터 2008년 7월까지 총 199개의 월별 유가로 이루어져 있다.

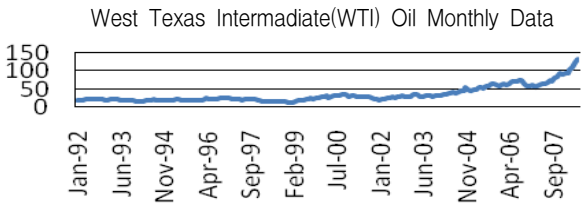

Figure 4. The monthly WTI crude oil prices from Jan. 1992 to Jun. 2008

동일기간에 대하여, 국제유가에 관련된 다양한 변수들이 고려 되었다. 초기 변수로는 에너지경제연구원에서 제공하는 국제원유시장에서 공급에 관련된 변수들, 수요에 관련된 변수들, 그리고 기타 외적인 경제지표들로서 총 103개의 변수들이 WTI 예측에 고려되었다. 이들 중, 본 연구에서는 이 분야 전문가의 자문을 통해 실무에서 WTI와 관련이 높다고 판단되는 변수들 26개를 사용하였다. 공급관련 변수로는 세계석유총생산량, OPEC 석유생산량, 사우디석유생산량 등이 있으며 수요관련 변수로는 세계석유총수요량, OECD 석유수요, NonOECD 석유수요 등이 있다. 기타 경제지표 관련변수로는 생산자물가지수, 미국달러환율 등이 있다. 다음의 <표 2>에서는 본 연구에서 사용한 26개 변수들을 보여준다. <그림 5>은 이러한 입력변수들이 예측하고자 하는 목표변수인 WTI와 어떻게 영향을

주고 받는지를 나타내고 있다. 특히, 입력변수들은 WTI에 영향을 미친다는 점 외에 입력변수들간에도 상호 영향을 주고 받는 복잡한 관계를 형성하고 있음을 알 수 있다.

3.2 입력변수 전처리 및 변환

26개의 입력변수 각각은 앞서 제 2.2절 <표 1>에서 소개한 바와 같이 7개의 기술적 지표들(TI) MA, BIAS, OSC, ROC, K, D, RSI로 변환되었다. <표 1>에서 각 TI별 파라미터 값으로는 $z \in \{3, 4, 6, 8, 9, 12\}$이 사용되었다. 예를 들어, 입력변수들 중 하나인 사우디석유생산량(SAUDI)은 MA 관련 파생변수 6개, BIAS 관련 파생변수 6개, OSC 관련 파생변수 6개, ROC 관련 파생변수 6개, K 관련 파생변수 6개, D 관련 파생변수 6개, RSI 관련 파생변수 6개로 총 42개의 변수로 지표화된다. 이와 같이, TI를 사용하면 시계열 데이터의 추세 및 기반구조를 고려할 수 있는 장점이 있는 대신에 하나의 변수가 기술적 지표별로 파라미터 값의 개수만큼 많은 파생변수들로 증가하게 되는 단점이 있다. 이 다수 개의 파생변수들을 모두 모델링에 사용할 경우 입력변수의 차수가 증가함으로 인해 예측 모델의 좋은 성능을 기대하기 어렵다. 따라서 이를 방지하기 앞서 제 2.3절에서 소개한 변수추출 및 차원감소 기법을 사용하였다. 즉 선형주성분분석(PCA) 또는 비선형주성분분석(NLPCA)을 사용하여 각 TI별 파생변수들로부터 1개, 3개, 6개의 특성지표(feature)들을 추출하였다. 예를 들어 특성변수를 1개로 설정한다면, 사우디 생산량(SAUDI)의 경우에는 총 42개의 파생변수들(= 기술적 지표수(7개) X 지표별 파라미터 값의 개수(6개))로 증가했다가 PCA 또는 NLPCA에 의해 다시 7개의 특성변수들로 차원감소된다. <그림 6>은 이와 같은 과정을 개략화 한

Table 2. The 26 sets of time series data

Input Variables	Target Variable
Overall amount of world oil demand, amount of OECD demand, non-OECD demand, China demand, USA demand	West Texas Intermediate (WTI) Crude Oil Prices
OPEC production, Saudi production, Iran production, Iraq production, Kuwait production, non-OPEC production, USA production, Russia production, world production	
producer price index, U.S. exchange rate, OECD commercial stockpiles, U.S. commercial stockpiles for crude oil, USA commercial stockpiles for oil, OPEC surplus production ability, NYMEX oil futures price, non-commercial real purchase(short), non commercial real purchase(long), commercial volume(short), commercial volume(long)	

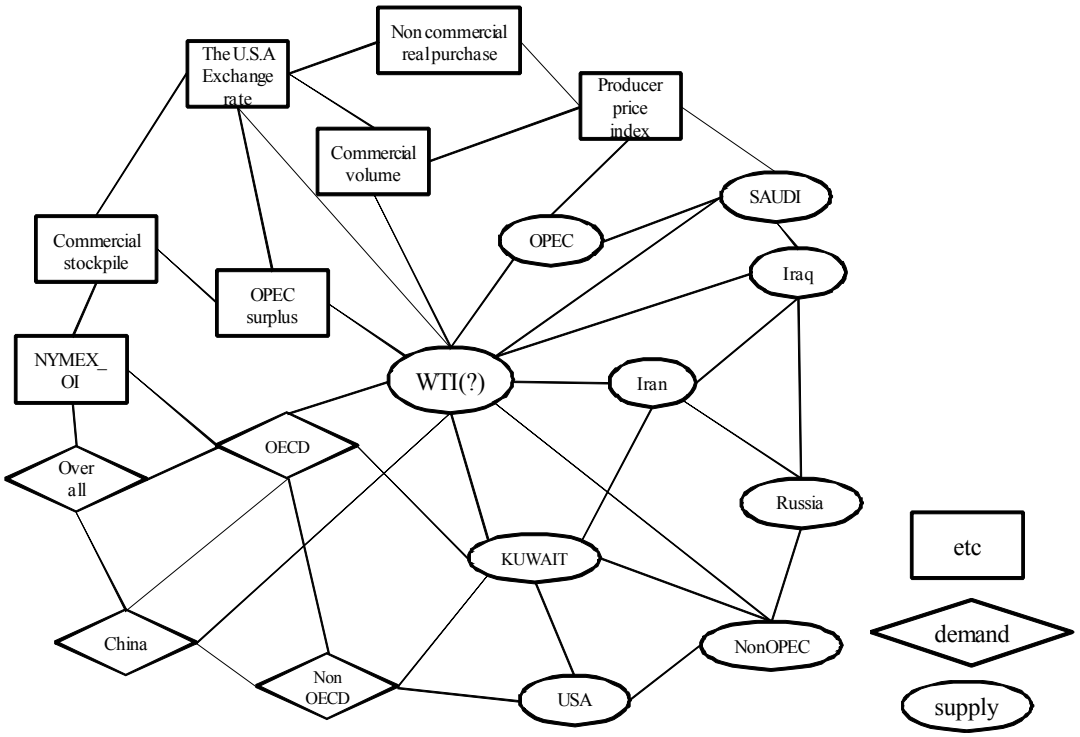

Figure 5. The SSL graph for the 26 sets of time series variable

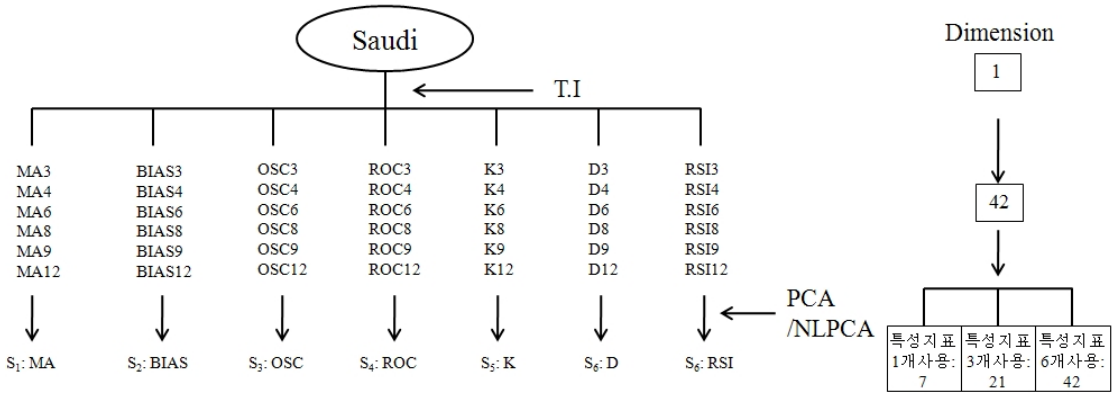

Figure 6. Feature extraction process for monthly productions of Saudi

것이다.

3.3 목표변수(y) 설계

SSL은 목표변수값이 있는 데이터 즉, 레이블 된 데이터와 그렇지 않은 데이터를 유사도 매트릭스에 의해 결합하여 결과값을 얻는 방법이다. 본 연구에서는 유가의 상승·하락에 대하여 다음의 식 (10)과 같이 레이블을 설정하였다.

$$y = \mathrm{sign}(x_t - MA_3(X_t)) \tag{10}$$

예를 들어, 이번 달의 SAUDI 석유생산량이 지난 석 달 동안의 이동평균보다 상승했으면 SAUDI의 이번 달 레이블은 가격이 올랐다는 의미로 +1, 감소했을 경우에는 -1로 설정한다. 만약 이러한 증가 및 감소의 정보가 주어지지 않을 경우에는 레이블은 0으로 설정되며 모델링의 결과로서 예측되어야 하는 미지수라는 의미이다. 본 연구에서는 유가예측을 위해 26개의 변수들 중 WTI 현물가격을 예측되어야 할 값으로 설정하고 나머지 25개의 변수에 식 (10)의 방법으로 레이블을 부여하였다.

3.4 SSL 유사도 매트릭스(Similarity Matrix)

SSL은 어떻게 유사도 매트릭스를 만드는가가 모델의 예측성능을 좌우할 수 있다(Shin *et al*., 2007; Shin *et al*., 2010). 유사도 매트릭스는 <그림 2>의 그래프에서 노드 간의 엣지에 대한 연결강도를 표현하며 그 값이 클수록 두 노드사이의 유사정도가 크다는 의미를 갖는다. 본 연구에서는 앞서 제 2장에서 기술한 바와 같이 각 노드마다의 시계열 데이터를 기술적지표화 및 변수추출과정으로 벡터화하고 이를 유사도 계산에 사용하였다. 즉, <그림 5>에서 각 노드들을 벡터들로 표현하고 <그림 6>의 과정을 따라 유사도 계산이 이루어졌다. 따라서 26개의 변수는 26개의 노드로 표현되며 노드간 연결은 유사도 매트릭스에 의해 결정되었다. 각 노드마다의 목표값(y)와 출력값(f)는 유사정도에 비례하여 이웃 연결된 노드와 상호 영향을 주고받게 된다.

3.5 SSL 예측값(f)에 대한 해석 방법

WTI에 대한 SSL 모델의 예측값은 제 2.1절의 식 (3)에 의해 f로 도출되며 이때 f의 값은 실수 값이다. 모델의 예측값(f)에 대한 해석방법은 다음과 같다. 예를 들어 현 시점이 t이고 모델의 WTI 예측값이 f = 0.124라면 이는 $X_t - MA_3(X_t) = 0.124$라는 의미이다. 이는 다음 달 WTI의 유가가 과거 3개월 이동평균보다 높게 상승한다는 의미가 된다. 따라서

$$\mathrm{sign}(X_t - MA_3(X_t)) > 0 \tag{11}$$

이므로 다음과 같은 부등식이 성립한다:

$$X_t > MA_3(X_t) \tag{12}$$

한편, <표 1>의 이동평균 MA의 공식

$$MA_3(X_t) = \frac{1}{3}(X_t) + \frac{3-1}{3}MA_k(X_{t-1})$$

과 식 (12)로부터 다음을 유도할 수 있다.

$$X_t > \frac{1}{3}X_t + \frac{2}{3}MA_3(X_{t-1})$$

따라서 결론적으로는 다음의 부등식을 얻는다.

$$X_t > MA_3(X_{t-1})$$

즉, f > 0이면 실제로 $X_t > MA_3(X_{t-1})$를 의미하므로 t-1시점에서 한 시점 앞선 t시점의 유가등락을 예측할 수 있다.

4. 결과

실험에 사용한 데이터 셋은 다음과 같이 설정하였다. 1993년 1월부터 2000년 5월까지의 100개의 월별 데이터를 학습데이터 셋(training set)으로, 2000년 6월부터 2008년 6월까지의 86개의

월별 데이터를 테스트셋(test set)으로 사용하였다(총 199개의 시점으로부터 기술적 지표생성에 필요한 13개의 시점이 제외되었으므로 총 186개의 벡터화된 데이터를 얻음). t 시점을 예측하기 위해서 SSL의 유사도매트릭스는 t-1시점에서 계산되었다. 일반적으로 SSL 모델은 학습데이터셋과 테스트셋을 나누는 의미가 크게 없으나 기존의 다른 모델들과의 비교를 위하여 구별하였다. 실험에서는 본 연구에서 제안한 7개의 SSL 모델들과 잘 알려진 대표성이 있는 5개의 모델들을 비교하였다. 7개의 SSL모델들은 각각 SSL_0, SSL_{P1}, SSL_{P3}, SSL_{P6}, SSL_{N1}, SSL_{N3}, SSL_{N6}로서, 변수추출방법으로 PCA를 썼는지 NLPCA를 썼는지, 추출된 특성변수의 개수가 몇 개인지에 따라 표기가 다르다. 예를 들어 SSL_{P3}은 기술적 지표로 변환한 후 PCA에 의하여 3개의 특성변수를 추출한 후 SSL을 적용한 모델을 의미한다. 마찬가지로 SSL_{N3}는 NLPCA로 3개의 특성변수를 추출하여 SSL에 적용한 모델이다. SSL_0로 표기된 모델은 기술적 지표화 후 변수추출은 생략한 모델을 말한다. 본 연구에서 비교용도로 사용한 5개의 모델들은 각각 자기회귀모형(auto-regression, AR), 로지스틱 회귀모형(logistic regression, LR), 인공신경망(artificial neural network, ANN), RBF 커널함수를 사용한 SVM(support vector machine, $SVM_{\_RBF}$), Polynomial커널함수를 사용한 SVM(support vector machine, $SVM_{\_POLY}$)이다. 각 모델들의 예측 정확도는 ROC 커브의 면적인 AUC(area under the ROC curve) 의 값으로 나타낸다. 본 연구에서 제안한 SSL 모델들의 파라미터 값은

$$\{k, \mu\} \in \{2, 3, 4, 5\} \times \{0.01, 0.1, 0.3, 0.5, 0.7, 1, 10, 100\}$$

로부터(k는 식 (1)의 k-nearest neighbors의 수, μ는 식 (2)의 loss-smoothness tradeoff를 의미) cross-validation을 통하여 최적의 조합으로 선정하였다. <그림 7>는 본 연구에서 제안한 7개의 모델들 중 가장 기본적인 모델인 SSL_0가 파라미터 k와 μ의 조합에 따라 AUC 성능이 어떻게 변화하는지를 예시한 것이다.

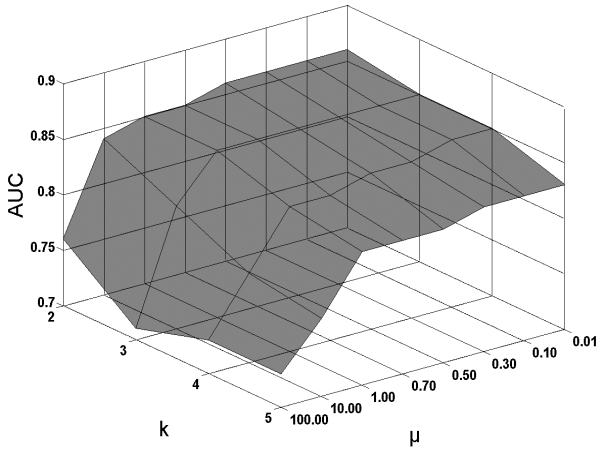

Figure 7. The AUC over parameters variation (k and μ) using model SSL_0

비교모델로 사용된 5개의 모델들 각각에 대하여도 유사한 방법으로 최적의 학습파라미터를 설정하였다. 다음의 <표 3>은 총 12개 모델들에 대한 AUC 결과값이다. <표 3>의 결과를 간략히 도표화 하면 다음 <그림 8>과 같다. 우선 시계열 예측에서 주로 사용하는 AR과 LR은 각각 0.53과 0.55의 평균 AUC를 보여 제안한 SSL 모델들의 평균인 0.82 AUC에 크게 못 미침을 알 수 있다. 이는 유가 예측이 기존의 선형 모델에 기반한 시계열모델로는 분석의 한계가 있음을 나타낸다. ANN과 SVM의 경우는 AUC 평균값이 0.74와 0.66으로 나타나 AR 및 LR 보다는 비교적 우수한 정확도를 보이지만 본 연구에서 제안한 SSL 모델들보다 상대적으로 저조한 결과를 나타내었다. 본 연구에서 제시한 SSL 모델들이 ANN과 SVM의 예측력보다 뛰어난 이유는 SSL의 경우 WTI 변수와 수요, 공급, 기타변수들 간의 일대일 관계뿐 아니라 그들간의 내재적 상호 관계까지 고려함으로써 더 정확한 예측이 가능하다고 분석된다.

Table 3. AUCs Comparison for different methods

		Max_AUC	Avg_AUC	Min_AUC
SSL	SSL_P	0.86	0.84	0.76
	SSL_{P1}	0.85	0.83	0.77
	SSL_{P3}	0.88	0.86	0.77
	SSL_{P6}	0.86	0.85	0.77
	SSL_{N1}	0.84	0.83	0.81
	SSL_{N3}	0.82	0.80	0.77
	SSL_{N6}	0.75	0.74	0.72
SSL Avg		0.84	0.82	0.77
비교 모델	AR	0.54	0.53	0.52
	LR	0.64	0.55	0.49
	ANN	0.82	0.74	0.55
	SVM_{RBF}	0.78	0.73	0.67
	SVM_{POLY}	0.74	0.58	0.50

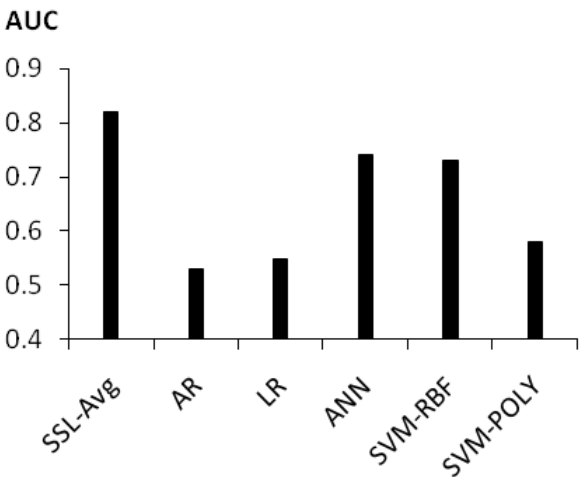

Figure 8. SSL vs. Others

다음의 <그림 9>에서는 본 논문에서 제안한 7개 SSL 모델들 간의 AUC값을 비교한다. TI만 사용하고 변수추출 기법을 사용하지 않은 SSL_0의 평균 AUC는 0.84로서 이미 다른 모델들(AR, LR, ANN, SVM)의 성능보다 우수하다. 이를 기준으로 제안한 7개의 모델들을 비교하면, PCA를 변수 추출기법으로 사용한 SSL_P모델들은 성능이 SSL_0와 유사하며, NLPCA를 변수추출기법으로 사용한 SSL_N모델들은 성능이 다소 감소하는 경향이 있다. 주목할 만한 점은 SSL_{P1}의 경우 7개의 특성변수들만으로도 42개의 변수를 사용한 SSL_0의 성능을 거의 회복했다는데 있다. 또한 SSL_{P3}의 경우에는 SSL_0보다 더 나은 성능을 나타낸다. SSL_N의 경우에는 SSL_{N1}과 SSL_{N3}의 성능은 SSL_0와 유사하나 SSL_{N6}에서는 오히려 성능이 감소하는 패턴을 보인다. 정리하면 변수추출기법을 사용한 경우 보다 적은 수의 변수들만으로도 원래의 예측 성능과 유사한 성능을 기대할 수 있으며 특히 본 실험에서는 NLPCA보다 PCA가 보다 효과적이라는 것을 알 수 있었다.

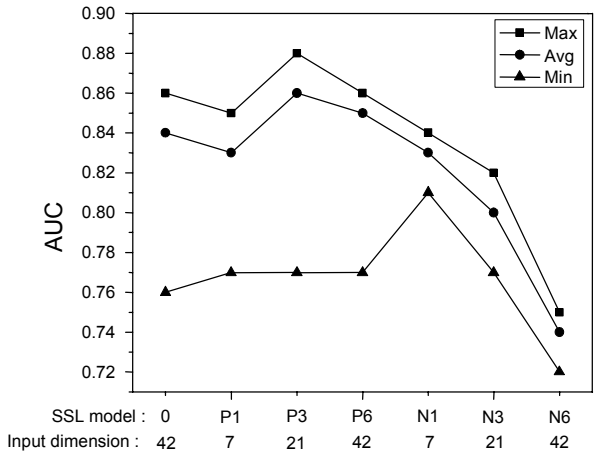

Figure 9. Comparison of AUC for Seven Different SSL Models

본 연구에서 제안한 SSL 모델들 중 변수의 수 대비 성능 측면에서 가장 우수하다고 사료되는 SSL_{P3} 모델과 비교 모델 중 성능이 우수하게 나온 $SVM_{\_RBF}$, ANN 모델들에 Confusion Matrix를 대입해 정확도를 비교하였다. <표 4>는 Confusion Matrix의 정확도를 계산한 방법이다. 정확도는 전체데이터에서 실제값과 예측값의 상승, 하락이 일치하는 부분의 백분율로 나타내었다

Table 4. Confusion Matrix

		Predicted Class		Accuracy, %
		Up	Down	
Actual Class	Up	a	b	$\frac{a+b}{a+b+c+d}\times 100\%$
	Down	c	d	

<표 5>에서 모델 별로 Confusion matrix 및 정확도를 나타내었다. <표 5>에서 보듯이 변수추출기법 및 그 개수가 적절히 선정 된다면, 제안한 SSL 모델기법이 다른 모델들에 비해 월등히 향상된 정확도를 보임을 알 수 있다.

Table 5. Confusion Matrix and Accuracy: SSL_{P3} vs. SVM_{RBF}, ANN >

predicted value (f) >= 0, Up predicted value (f) < 0, Down			predicted value Up	predicted value Down	Accuracy, %
Model	SSL_{P3}	Up	55	7	86.04%
		Down	5	19	
	SVM_{RBF}	Up	41	21	72.09%
		Down	3	21	
	ANN	Up	37	25	67.44%
		Down	3	21	

5. 결론

본 논문은 유가예측 분야에서 다양한 최신기계학습기법을 도입하여 새로운 방법론을 소개하고 예측정확도를 향상시켰다는데 의의를 둘 수 있다. 본 논문에서 제안하는 방법은 다음과 같은 이점이 있다. 방법론적 측면에서는 유가의 등락에 변동을 미치는 수요, 공급 및 국제 경제지표 같은 시계열 데이터들의 직·간접적인 인과관계를 SSL을 변형 활용하여 모델링하였다. 기술적으로 특히 SSL을 사용함으로써 독립변수가 종속변수에 미치는 영향뿐만 아니라 독립변수간의 상호영향도 모델링에 포함되게 하여 좀 더 정확한 유가예측을 가능하게 하였다. 이는 기존의 유가의 시계열적 특성만을 고려하여 예측하는 연구와는 색다른 접근 방법이다. 세부적으로는 인과관계를 표현하기 위해서 시계열 데이터끼리의 유사성을 계산해야 하는데 시계열 데이터자체의 잡음을 해결하기 위해 기술적 지표화(T.I)함으로써 시계열 데이터의 잡음을 제거하고 동시에 저변에 존재하는 추세 및 구조적 변동요인을 반영하게 할 수 있다. T.I 도입시 기술적인 문제점으로 차원이 증가하는 문제점이 있는데, 변수변환 및 차원감소기법을 사용함으로써 변수들간 공통적으로 내재된 특성변수를 추출하였고 이는 입력변수의 불필요한 증가를 지양하게 하는 효과를 거두었다. 본 연구에서는 위의 기법들이 조합된 시너지 효과로서 평균 AUC 0.86, 정확도 86%라는 매우 우수한 예측정확도를 도출하였으며, 이는 시계열 분석에서 좀처럼 얻기 어려운 우수한 성능으로 사료된다. 본 연구에서 제안한 방법론은 예측이 필요한 모든 시계열 분석에 적용될 수 있다. 앞서 예로 제시한 바와 같이 국제유가 변동, 국내외 주가지수 흐름 예측, 물가 변동성 예측, 국가 성장률 예측, 환율예측 등의 여러 분야에 확장, 적용될 수 있으리라 기대한다.

참고문헌

Abosedra, S. and Baghestan, H. (2004), On the predictive accuracy of crude oil prices, *Energy Policy*, **32**, 1389-1393.

Akarca, A. T. and Andrianacos D. (1997), Detecting break in oil price series using the box-tiao method, *International Advances in Economic Research*, **3**(2), 217-224.

Amano, R. A. and Norden, S. V. (1998), Exchange rates and oil prices, *Review of International Economics*, **6**(4), 683-694.

Basher, S. A. and Sadorsky P. (2006), Oil price risk and emerging stock markets, *Global Finance Journal*, **17**, 224-251.

Birol, F. (2004), Analysis of the impact of high oil price on the global economy, *International Energy Agency*, 1-15.

Cheng, H. and Tan, P. N. (2008), Semi-supervised Learning with Data Calibration for Long-Term Time Series Forecasting, *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 133-141.

Cortazar, G. and Schwartz, E. S. (2003), Implementing a stochastic model for oil futures prices, *Energy Economics*, **25**, 215-238.

He, L. Y., Fan, Y., and Wei Y. M. (2009), Impact of speculator's expectations of returns and time scales of investment on crude oil price behaviors, *Energy Economics*, **31**, 77-84.

Keong, Y. S. (2009), Economic forecasting-de-mystifying the art of modern crystallomancy, *Singapore Institute of Statistics*.

Lanza, A., Manera, M., and Giovannini, M. (2005), Modeling and forecasting cointegrated relationships among heavy oil and product prices, *Energy Economics*, **27**, 831-848.

Mirmirani, S. and Li, H. C. (2004), A comparison of VAR and neural networks with genetic algorithm in forecasting price of oil, *Advances in Economics*, **19**, 203-223.

Morana, C. (2001), A semiparametric approach to short-term oil price forecasting, *Energy Economics*, **23**, 325-337.

Shin, H., Lisewski, A. M., and Lichtarge, O. (2007), Graph sharpening plus graph integration: a synergy that improves protein functional classification, *Bioinformatics*, Oxford University Press, **23**(23), 3217-3224.

Shin, H., Hill, N. J., Lisewski, N. J., and Park J. S. (2010), Graph sharpening, *Expert Systems with Applications*, **37**, 7870-7879.

Stevens, P. (1995), The determination of oil prices 1945～1995, *Energy Policy*, **23**(10), 861-870.

Svensson, L. E. O. (2005), Oil prices and ECB monetary policy, *Committee on Economic and Monetary Affairs*, 1-4.

Verleger, P. K. (1993), Adjusting to volatile energy prices, *Institute for International Economics*, **23**(3), 325-338.

Xie, W., Yu, L., Xu, S. Y., and Wang, S. Y. (2006), A new method for crude oil price forecasting based on support vector machines, *International Conference on Computational Science*, **3994**, 444- 451.

Yousefi, S., Weinreich, I., and Reinarz, D. (2005), Wavelet-based prediction of oil prices, *Chaos Solitons and Fractals*, **25**, 265-275.

Yu, L., Wang, S. Y., and Lai, K. K. (2008), Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm, *Energy Economics*, **30**(5), 2623-2635.