VFAST Transactions on Software Engineering
http://vfast.org/journals/index.php/VTSE@ 2014 ISSN(e): 2309-3978;ISSN(p): 2411-6246
Volume 2, Number 1, January-December, 2014

pp. 21-28

A REVIEW OF PROBABILISTIC GRAPH MODELS FOR FEATURE SELECTION WITH APPLICATIONS IN ECONOMIC AND FINANCIAL TIME SERIES FORECASTING

TAHSEEN AHMED JILANI[1], AND SYED ALI RAZA NAQVI[2]
[1]Department of Computer Science, University of Karacchi
tahseenjilani@uok.edu.pk
[2]Department of Statistics, University of Karacchi
naqvi.stats@gmail.com

Revised March 2014

ABSTRACT. *In every field of life, people are interested to be able to forecast future. A number of techniques are available to predict and forecasting upto a certain level of accuracy. Many techniques involve statistical tools and techniques for forecasting, modeling and control. Use of statistical techniques is growing with time and new techniques are being developed very rapidly. Especially in the field of economics and finance, the estimation and forecasting of economic and financial indicators play a vital role in decision making. Many models are developed in the last 2 decades to get better accuracy and efficiency in time series analysis and still there is a scope of learning and getting betterment in this field is available. In this research we have reviewed probability graphs, directed acyclic graphs, Bayesian networks, feature selection algorithms and Markov blankets for time series forecasting on the economic and financial problems (like stock exchange forecasting, multi-objective business risk analysis, consumers' analysis, portfolio optimization, credit scoring etc). This is a new dimension for adaptive modeling techniques in economics and finance modeling.*

Keywords: Features selection (FS); Probabilistic Graph Models (PGMs); Bayesian networks (BN); Markove blanket (MB); Autoregressive Integrated Moving Average models (ARIMA) and Generalized Autoregressive Conditional Heteroskedasticity models (GARCH).

1. **Directed acyclic graphs.** Directed acyclic graphs (DAGs) are commonly used tool for formulating and designing the statistical, logical and mathematical problems based on graph theory. DAGs make the visualization of the interaction of the variables easy and more meaningful. In DAGs, each end represented by a circle shows the variable and the connected arrow line shows the effect of one variable on the other. DAGs play a vital role in feature selection for any statistical study. Currently, step wise regression and model selection criterion are commonly used for this purpose to enter or remove a variable in a model which becomes difficult in some specific situations.

Graph representations have gained quite some popularity in the past few years. They offer a strong paradigm in terms of representational power mainly thanks to their ability to encode relations among the elements of a given pattern. Graphs have been extensively used in regression modeling, Bayesian inference, epidemiology, bioinformatics, link analysis, computer network analysis, web content mining and in many other subfields of statistics and computer science [8]. Directed acyclic graphs (DAGs) are graphical representations of the assumed causal relations between the exposure of interest, the outcome, and key covariates [6]. A directed acyclic graph (DAG) can describe parallel computations where nodes represent tasks and directed arcs

represent synchronization constraints among tasks. A DAG may be analyzed for mean completion time, for time of the shortest path and for various other performance measures [4]. Importance of DAGs in representing independencies, conditional independencies and causal relationships by proposeing a decomposition approach for structural learning of DAGs. In this approach, a problem of learning a large DAG is split into problems of learning small sub-graphs. Domain or prior knowledge of conditional independencies can be utilized to facilitate the decomposition of structural learning. They theoretically proved the correctness of the proposed algorithms. Both the complexity of the algorithms and the power of conditional independence tests can be improved by decomposing a large graph into small sub-graphs. The theoretical results can also be used for scheme design of multiple datasets [32].

DAGs with no directed cycles is a fresh method for visualizing finite DAGs in such a way that the nodes do not overlap. Finally they applied their research on a new E-spring algorithm for removing node overlaps in clustered directed acyclic graphs (DAGs). The overall graph area reduction achieved using enhancements to E-Spring algorithm varies between 45% and 79%, depending on the string length and number of clusters [22].

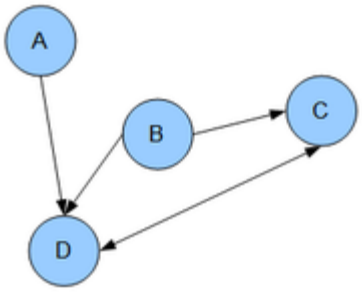

Figure 1. Directed Acyclic Graph

DAGs are widely used nowadays in reactive scheduling of heterogeneous and dynamic computing systems, rescheduling strategies for grid workflow, multiclass classification, modeling event driven, genes recognition, RNA and DNA identification and prediction, Variation recording, Taxonomy, Training and development, Sequence analysis, Protein classification and clustering, mapping, Homology, Literature, genetics and medicines, data and software designing and manipulation and chemical and bioassay fields of life.

2. **Probabilistic Graph Models.** Probabilistic graphical models (PGMs) combine ideas from statistics and computer science into a unifying framework for modeling complex real-world phenomena. PGMs are now widespread in language and speech processing. Probabilistic graphical models are a combination of probability and graph theories. They provide a natural tool for dealing with two problems uncertainty and complexity that occur throughout applied mathematics, applied statistics and engineering especially and in particular they are playing an increasingly important role in the design and analysis of machine learning algorithms. Fundamental idea of a PGM is conceived from a complex system which is built by combining simpler parts. Probability theory provides the glue whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data. The graph theoretic side of graphical models provides an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure [19]. Graphical modeling (GM) is a relatively new statistical approach that earned famous in the last decade. The major attraction of approach in empirical research is its convenient way to present pairwise relationships between random variables taken from a multivariate context. The initial step in the approach is the computation of the partial correlations between the variables in the particular multivariate system under study. Once the numerical values are known we can test their significance by using an opportune statistic. Finally the results are presented as a graph. If the variables in the graph are jointly distributed as a multivariate Gaussian distribution, a significant partial correlation implies the presence of conditional dependence. For this reason the graph is called a conditional independence graph (CIG) [24]. A more informative object in GM is the directed acyclic graph (DAG). This is a directed graph where there are arrows linking the nodes and where the joint distribution of the variables can be expressed as a sequence of marginal conditional distributions [24].

PGMs talk over a mixed bag of models, spreading over discrete and continuous Bayesian systems, undirected Markov systems, and developments to manage dynamical and relational frameworks. These models can additionally be taken in immediately from data, permitting the methodology to be utilized within situations where physically building a model is challenging or even impossible. The framework of the PGMs provides algorithms for discovering and analyzing structure in complex distributions to describe them in brief and extract the unstructured information, allows them to be built and utilized effectively [21].

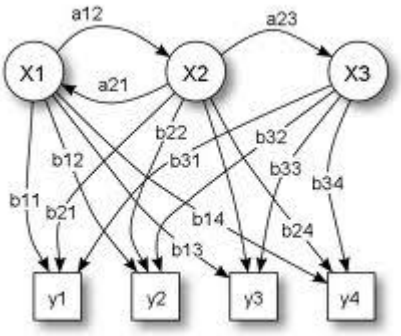

Figure 2. Probabilistic Graph Model

Applications of PGMs includes many of the classical multivariate probabilistic systems studied in different fields such as pure statistics, statistical mechanics, systems engineering, information theory and pattern recognition. PGMs can also be included in mixture models, factor analysis, hidden Markov models and these are well suited to handle the problems regarding complex and structured relationships, a large number of relevant attributes, and large volumes of data.

3. Feature Selection. In statistics, feature selection, otherwise called variable selection which is the procedure of selecting a subset of significant variables for utilization in model development. The focal surmise when utilizing a feature determination system is that the information holds numerous repetitive or insignificant features. Feature choice procedures give three fundamental profits when developing prescient models:

- improved model interpretability,
- shorter training/learning times,
- enhanced generalization by reducing over fitting.

Feature selection is defined by many authors as an ideal tool that may be used for model's redundancy elimination, resolving time and model complexity reduction, improving prediction accuracy and approximating actual class distribution. Feature selection attempts to select the minimally sized subset of features according to the following criteria. The criteria can be:

1. The classification accuracy does not significantly decrease; and
2. The resulting class distribution, given only the values for the selected features, is as close as possible to the original class distribution, given all features.

The most popular form of feature selection used in statistics is stepwise regression. It is a greedy algorithm that adds the best feature (or deletes the worst feature) at each round. The main control issue is deciding when to stop the algorithm. There are a variety of optimality criteria that can be used for controlling feature selection including, C_p statistic, Akaike information criterion (AIC), Bayesian information criterion (BIC), minimum description length (MDL), and a variety of new criteria that are motivated by false discovery rate (FDR) [28].

The feature selection process constitutes a commonly encountered problem of global combinatorial optimization. This process reduces the number of features by removing irrelevant, noisy, and redundant data, thus resulting in acceptable classification accuracy. Feature selection is a preprocessing technique with great importance in the fields of statistics, data analysis and information retrieval processing, pattern classification, and data mining applications [3].

Feature selection is the process of defining the most informative and discriminative features in a dataset for

the data mining task. The two basic steps in a typical feature subset selection process are the specification of the parameter set and the search for the best subset. The parameter set often includes the selection algorithm, the learning mechanism, and the process for error estimation. It is commonly observed that there is no single best parameter set combination valid for all tasks and all types of datasets. In addition, the performance of the feature selection process is strongly dependent on the selection algorithm employed [30].

The presence of less relevant or highly correlated features often decreases classification accuracy. Feature selection in which most informative variables are selected for model generation is an important step in data-driven modeling to get discriminating power and model performance [31].

4. **Bayesian Networks.** A Bayesian network is a directed acyclic graph in which the nodes represent the variables and the arcs represent a relationship among the connected variables. A conditional probability table gives the strength of such relationship. The variables that are not connected by an arc can be considered as having no direct influence on them. The strength of this influence is given by the conditional probability

$$p(x) = \prod_{v \epsilon V} p(x_v / x_{pv(v)})$$

These conditional probabilities in the graph are often estimated by using known statistical and computational methods. Hence, BNs combine principles from graph theory, probability theory, computer science, Mathematics and Statistics.

The most specific use of DAG is a probabilistic graphs models (PGMs). In PGMs, DAGs are used to identify the conditional independence among the variables (also by using probability theories) which make development of a statistical model more easy, accurate and reliable. PGMs are always proven efficient for probabilistic inference like Bayesian networks (BNs). They are the directed acyclic graphs which are used to compute the conditional probability among the variables [21].

A Bayesian network graphically encodes probabilistic relationships among a set of features. It has numerous applications from diagnosis and forecasting to automated vision and manufacturing control [13]. A Bayesian network can be constructed on the basis of a multivariate approach called causal structure learning [25] this technique aims to discover causal relationship among variables and to build a directed acyclic graph (DAG) from observational data. Constructing a DAG is a difficult problem with exponential complexity in the worst case. [1].

Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph. The use of the Bayesian network enables the time series forecast without white noise model. Although the time series data are continuous values, the Bayesian network can deal with the discrete (digitized) values alone. The time series distribution is digitized firstly by using the clustering algorithms and then, the Bayesian network is used for modeling the stochastic dependencies among the digitized values of the previous time series [33].

Bayesian networks (BN) are statistical models which allow graphic representation by means of a directed acyclic graph (DAG), complex structures of dependencies among stochastic variables. They have been widely employed in a great variety of real world problems because of their excellent properties for reasoning under uncertainty [21].A Bayesian Network is ultimately a graphical representation of a collection of conditional independence restrictions [5]. Bayesian networks play an important role in decision making and statistical inference and have been applied in many fields. They offer powerful knowledge representations for independence, conditional independence and causal relationships among variables in a given domain [14].

Bayesian Network is popular technique due to their ability to support probabilistic reasoning from data with uncertainty. According to the BNs, probabilistic inference can be conducted to predict the values of some variables based on the observed values of other variables. Hence, we can conclude four principal advantages of using a discrete variable Bayesian network for network analysis: (1) the Bayesian network framework does

not require that the joint distribution follows a specific parametric distribution; (2) a Bayesian network supports probabilistic reasoning as it consists of probabilistic associations among variables; (3) because the Bayesian network representation is based on the concept of conditional independence, it supports Bayesian inference without having to maintain the full joint distribution in memory; and (4) since each Bayesian network is a multivariate model that we can evaluate using a single probability score, we can evaluate many structure–function interactions without the multiple comparison problem [29].

Bayesian networks, equivalently graphical Markov models determined by directed acyclic digraphs or DAGs, have proved to be both effective and efficient for representing complex multivariate dependence structures in terms of local relations. However, model search and selection is potentially complicated by the many-to-one correspondence between DAGs and the statistical models that they represent [9].

The conditional probabilities in the BNs are mostly estimated by using statistical and computational_methods which then are used to develop a Markov blanket of the target variable. A Markove blanket is a Bayesian network of the target variable surrounded by its directly related variables called its children, parents and spouse. The Markov blanket holds all the variables that shield the target variable from whatever remains of the system. This technique is quite useful to reduce the complexity of the statistical model development.

5. **MARKOV BLANKETS**. Markov blanket filtering is a backward feature selection algorithm. Initially, for each feature in feature subset a Markov blanket is defined. The Markov blanket M_i for feature f_i is defined as the S features that have the highest Pearson correlation with feature f_i, where S is the size of the Markov blanket. For each feature, the coverage of its blanket (i.e. how well the information in a feature is covered by its blanket) is computed. The feature that has the lowest score (i.e. the highest coverage) is considered to be the most redundant and is removed from the dataset. This process is iterated until the feature subset contains S features. This is a logical stopping criterion as at this point it is no longer possible to create blankets of size S [18].

Markov blanket for any variable X, MB(X) is a minimal set of variables conditioned by which X is conditionally independent of all the remaining variables. Under the faithfulness assumption, ensuring that all the conditional independencies in the data distribution are strictly those entailed by G, MB(X) consists of the union of the set of parents, children, and parents of children (i.e., spouses) of X [25].

The Markov blanket for a node A in a Bayesian network is the situated of nodes ∂A made of A's parents, its children, and its children's other parents. In a Markov network, the Markov blanket of a node is its set of neighboring nodes. A Markov blanket may also be denoted by MB(A).

Every set of nodes in the network is conditionally independent of A when conditioned on the set ∂A, that is, when conditioned on the Markov blanket of the node A. The probability has the Markov property; formally, for distinct nodes A and B

$$P(A/\partial A, B) = P(A/\partial A)$$

The Markov blanket holds all the variables that shield the nodes from whatever remains of the system. This implies that the Markov blanket of a node is the main learning required anticipating the conduct of that node. [26]. In a Bayesian network, the qualities of the parents and kids of a node obviously give data about that node; nonetheless, its youngsters' guardians likewise must be incorporated, on the grounds that they might be utilized to clarify away the hub being referred to. The Markov Blanket studying calculation is a directed calculation that is utilized to discover a Bayesian Network that portrays the Target hub. All Discriminative models though the Markov Blanket Learning calculation gives back a Generative model.

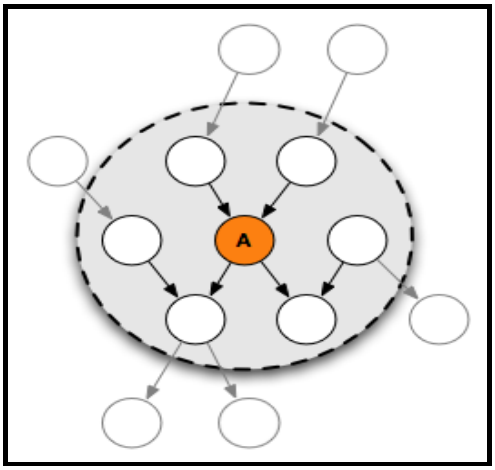

Figure 3. Markov Blanket

6. **Time Series Models.** The combination of autoregressive and moving averages models is known as autoregressive–moving-average (ARMA) models, which provide a parsimonious description of a stationary stochastic process in terms of two polynomials, one for the auto-regression and the second for the moving average. The notation ARMA (p, q) refers to the model with p autoregressive terms and q moving-average terms. This model contains the AR (p) and MA (q) models,[2]

In statistics and econometrics, and specifically in time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. ARMA models are fit to time series data either to better realize the data or to anticipate future focuses in the series (to get the expected values) [6]. They are connected in a few situations where data indicate confirmation of non-stationary, where a starting differencing step (relating to the "differencing" part of the model) might be connected to remove the non-stationary. If an autoregressive moving average model (ARMA model) is assumed for the error variance, the model is a generalized autoregressive conditional heteroskedasticity (GARCH) model. Usually, when testing for heteroskedasticity in econometric models, the best test is the White test. On the other hand, when managing time arrangement information, this intends to test for ARCH and GARCH blunders [10].

7. **Challenges in Probabilistic Graph Models for Time series forecasting:** This research direction is new and revolutionary to resolve the forecasting problems in the field of time series forecasting in terms of increasing forecasting accuracy (minimize the forecasting error), reducing the model complexity by reducing the variables by eliminating the non-useful variables based on feature selection techniques (only those variables will included who becomes in the Markov blanket) and applying and testing the conditional independence among the selected features for good forecasting. It is possible to use these techniques and develop better forecasting models in every field for better accuracy and minimum forecasting errors.

REFERENCES

[1] Bui, A. T., & Jun, C. H. (2012). Learning Bayesian Network Structure Using Markov Blanket decomposition. *Pattern Recognition Letters (33)*, pp. 2134 - 2140

[2] Box, G., Jenkins, G. M. & Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control (Third ed.). Prentice-Hall India*.

[3] Chuang, L. Y., Tsai, S. W., & Yang, C. H. (2011). Improved Binary Particle Swarm Optimization Using Catfish Effect for Feature Selection. *Expert Systems with Applications (38)*, pp. 12699 - 12707

[4] Colajanni, M., Presti, F. L. & Tucci, S. (2000). A Hierarchical Approach for Bounding the Completion Time Distribution of Stochastic Task Graphs, *Performance Evaluation (41)*, pp. 1 - 22

[5] Eckel, C. C., Gamalb, M. A. E. & Wilson, R.K. (2009). Risk Loving After the Storm: A Bayesian-Network Study of Hurricane Katrina Evacuees. *Journal of Economic Behavior & Organization (69),* pp. 110 - 124

[6] Engle, Robert F. (2001). The Use of ARCH/GARCH Models in Applied Econometrics. *Journal of Economic Perspectives 15(4),* pp.157 - 168

[7] Flander, W. D., Johnson, C. Y., Howards, P. P. & Greenland,S. (2011). Dependence of Confounding on the Target Population: *A Modification of Causal Graphs to Account for Co-Action (21),* pp. 698 - 705

[8] Gibert, J., Valveny, E. & Bunke, H. (2012). Feature Selection on Node Statistics Based Embedding of Graphs. *Pattern Recognition Letters (33),* pp. 1980 - 1990

[9] Gillispie, S. B. & Perlman, M. D. (2002). The Size Distribution for Markov Equivalence Classes of Acyclic Digraph Models. *Artificial Intelligence (141),* pp. 137 - 155

[10] Gujarati, D. N. (2003), *Basic Econometrics, Prentice-Hall India*.

[11] H. Saima, J. Jaafar, B.S. Belhaouari & Jillani T. A. (2011). ARIMA based Interval Type-2 Fuzzy Model for Forecasting. *International Journal of Computer Applications 28(3),* pp. 17 - 21

[12] Hacker, R. S. & Hatemi, J. A. (2005), A Test for Multivariate ARCH Effects. *Applied Economics Letters 12(7),* pp. 411 - 417

[13] Heckerman, D., Geiger, D. (1997). A Characterization of the Dirichlet Distribution Through Global and Local Parameter Independence. *The Annals of Statistics (25),* pp. 1344 - 1369.

[14] Jensen, F. V., Nielsen, T. D. (2007). Bayesian Networks and Decision Graphs. *Information Science and Statistics series (2nd ed.), Springer, New York-USA*

[15] Jilani T. A. , Burney S. M. A. & Ardil C. (2007). Multivariate High Order Fuzzy Time Series Forecasting for Car Road Accidents. *International Journal of Computational Intelligence 4(1),* pp. 15 - 20

[16] Jilani T. A. & Burney S. M. A. (2008). A Refined Fuzzy Time Series Forecasting Model For Stock Market Forecasting, *Physica A- vol. 387 (Statistical Mechanics with Applications),* pp. 2857 - 2862

[17] Jilani T. A., Mastorakis N., & Amjad U.(2012). A Hybrid Genetic Algorithm and Particle Swarm Optimization Based Fuzzy Time Series Model TAIFEX and KSE-100 Forecasting. *NAUN-First International Confernce on Biological Inspired Computation, University of Algaro, Portugal, 22-24, May-2012*

[18] Jilani T. A. & Burney S. M. A. (2008). Multivariate Stochastic Fuzzy Forecasting Models, *Expert Systems With Applications (ESWA),*(vol. 35, pp. 691 - 700)

[19] Jørgen, B.J. (2008), Acyclic Digraphs, Digraphs Theory, Algorithms and Applications, *Springer Monographs in Mathematics (2nd ed.), Springer-Verlag-Germany.*

[20] Knijnenburg, T. A., Reinders, M. J. T. & Wessels, L. F. A. (2006). Artifacts of Markov Blanket Filtering Based on Discretized Features in Small Sample Size Applications. *Pattern Recognition Letters (27),* pp. 709 - 714

[21] Koller, D.; Friedman, N. (2009). *Probabilistic Graphical Models, Massachusetts, MIT Press USA.*

[22] Kumar, P. & Zhang, K. (2009). Node Overlap Removal in Clustered Directed Acyclic Graphs. *Journal of Visual Languages and Computing (20),* pp. 403 - 419

[23] Masegosa, A. R. & Moral, S. (2013). New Skeleton Based Approaches for Bayesian Structure Learning of Bayesian Networks. *Applied Soft Computing (13),* pp. 1110 - 1120

[24] Oxley, L., Reale, M. & Wilson, G. T. (2008). Constructing Structural VAR Models with Conditional Independence Graphs. *Mathematics and Computers in Simulation (79),* pp. 2910 - 2916

[25] Pearl, J. (2000). *Causality. Cambridge University Press, Cambridge.*

[26] Pearl, J. (2010). On the Consistency Rule in Causal Inference (Axiom, Definition, Assumption or Theorem), *Epidemiology (21),* pp. 872 - 875

[27] Pearl, J., (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufman Publishers, San Mateo, CA.*

[28] Peng, H. & Ding, C. (2005). Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *Journal of Bioinformatics and Computational (3),* pp. 185 - 205

[29] Sun, Y., Tang, Y., Ding, S., Lv, S. & Cui, Y. (2011). Diagnose the Mild Cognitive Impairment by Constructing Bayesian Network with Missing data. *Expert Systems with Applications (38),* pp. 442 - 449

[30] Unler, A., Murat, A. & Chinnam, R. B. (2011). A Maximum Relevance Minimum Redundancy Feature Selection Method Based on Swarm Intelligence for Support Vector Machine Classification. *Information Sciences (181)*, pp. 4625 - 4641

[31] Vieiraa, S. M., Sousaa, J. M. C. & Kaymak, U. (2012). Fuzzy Criteria for Feature Selection, *Fuzzy Sets and Systems (189)*, pp. 1-18

[32] Xie, X., Geng, Z. & Zhaoa, Q. (2006). Decomposition of Structural Learning about Directed Acyclic Graphs. *Artificial Intelligence (170)*, pp. 422 - 439

[33] Zuo, Y. & Kita, E. (2012). Stock Price Forecast Using Bayesian Network, *Expert Systems with Applications (39)*, pp. 6729 - 6737.