A Novel Agent-DKGBM Predictor for Business Intelligence and Analytics toward Enterprise Data Discovery

Samaher Al_Janabi
Department of Information Networks, Faculty of Information Technology
University of Babylon,
samaher@itnet.uobabylon.edu.iq.

Abstract:

Today's business environment requires its workers to be skilled and knowledgeable in more than one area to compete. Data scientists are expected to be polyglots who understand math, code and can speak the language of business. This paper aims to develop a new Agent- Develop Kernel Gradient Boosting Machine (Agent-DKGBM) algorithm for prediction in huge and complex business databases. The Agent-DKGBM algorithm executes in two phases. In the first phase, building the cognitive agent, the primary goal is to prepare the database for the second phase, searching the business databases. During this phase, the cognitive agent selects one of the business databases, choosing the most suitable type i.e., Hyperbolic functions, Polynomial functions and Gaussian mixture as a kernel of Develop Support Vector Regression (DSVR) and determines the optimum parameters of the DSVR and DKGBM. The second phase consists of three stages, which include splitting the business databases into training and testing datasets by using 10-fold cross validation. In the second stage, a DKGBM model using the training data set is built to replace the Gradient Boosting Machine (GBM) kernel, typically using Decision Trees (DTs) to produce the predictor with DSVR because it would potentially increase the accuracy and reduce the execution time in the DKGBM model. Finally, the DKGBM would be verified based on the testing data set. Experimental results indicate that the proposed Agent-DKGBM algorithm will provide effective prediction with a significant high level of accuracy and compression ratio of execution time compared to other prediction techniques including CART, MARS, Random Forest, Tree Net, GBM and SVM. The results also reveal that by using Gaussian mixture as a kernel of DSVR, the Agent-DKGBM achieves more accurate and better prediction results than other kernel functions, which prediction algorithms typically use, also than GBM, which typically use DTs. Results clearly show that the proposed Agent-DKGBM improves the predictive accuracy, speed and cost of prediction. In addition, the results prove that Agent-DKGBM can serve as a promising choice for current prediction techniques.

Keywords: Agent, Business sectors, Agent-DKGBM, Develop Support Vector Regression (DSVR), Gradient Boosting Machine (GBM), Prediction Techniques, Smart Data.

الخلاصة

تتطلب بيئة الاعمال اليوم من عمالها ان يكونوا اكثر دراية ومهارة في اكثر من مجال تنافس. وعلماء البيانات من المتوقع ان يكونوا هم الاشخاص الذين يفهمون بالرياضيات والكودات ويتحدثون بلغة الاعمال. هذا البحث يهدف الى تطوير خوارزمية تنبى جديدة تدعى Agent-DKGBM *للتنبيء بقواعد بيانات كبيرة ومعقدة خاصة بالاعمال.* تنفذ هذه الخوارزمية Agent-DKGBM على طورين. في الطور الأولى، يتم بناء وكيل معرفي، والهدف الأساسي منه هو تحضير قاعدة بيانات للطور الثانية، خلال هذه الطور، الوكيل المعرفي يختار إحدى قواعد البيانات الاعمال، وكذلك يختيار افضل نوع من الدوال كنواة كـ *Hyperbolic functions, Polynomial functions and Gaussian mixture* لتطوير المتجهات الانحدار (DSVR)، ويحدد المعاملات المثلى لـ DSVR وDKGBM . ويتكون الطور الثانية من ثلاث مراحل، والتي تشمل تقسيم قواعد بيانات الاعمال الى قواعد بيانات التدريب والاختبار باستخدام باستخدام *10-fold cross validation*. في المرحلة الثانية، تم استخدام مجموعة بيانات التدريب لبناء DKGBM بعد ان تم استبدال النواة الاساسية لل GBM وهي اشجار القرار الثنائية بالـ DSVR والتي من المؤمل انها سوف تزيد من دقة وتقليل وقت التنفيذ في نموذج DKGBM. وأخيرا، سيتم التحقق من DKGBM استنادا إلى مجموعة بيانات الاختبار. النتائج التجريبية تشير إلى أن خوارزمية Agent-DKGBM المقترح تجهزنا بتنبيء فعال مع مستوى عال من الدقة وتقلل من وقت التنفيذ بالمقارنة مع غيرها من تقنيات التنبؤ المتضمنة CART, MARS, Random Forest, Tree Net, GBM and SVM. وتشير النتائج أيضا أن استخدام Gaussian mixture كنواة لـ DSVR ، وكذلك استبدال DTs الموجودة في GBM بالـDSVR يجعل الى Agent-DKGBM تحقق نتائج تنبؤ أكثر دقة وأفضل من استخدام الدوال الاخرة كنواة تظهر النتائج بوضوح أن Agent-DKGBM يحسن كل من دقة والسرعة والتكلفة للتنبؤ. وبالإضافة إلى ذلك، فإن النتائج تثبت أن Agent-DKGBM يمكن أن تكون بمثابة خيار واعد لتقنيات التنبؤ الحالية.

1. Introduction

Business sector based on data mining, which is capable for discovering hidden patterns of sales and market has been one of the most popular and widely used tools for identifying business choices in sales and marketing of new products or deriving market business future-

decisions. The main challenges in business environments for solving any problem can summarize by the following points: (i) Learn the business's processes and the data that is generated and saved, (ii) Learn how people are handling the problem now and what metrics they use or ignore to gauge success. (iii) Solve the correct, yet often misrepresented,problem using the optimization model, mathematical model and novel model, (iv) Learn how to communicate the above effectively (John, 2014).

Traditional data analysis techniques focus on mining quantitative and statistical data. These techniques aid useful data interpretations and help provide understanding of the practices deriving the data. While these techniques are helpful, they are not automated and can be prone to error as they rely upon human intervention by an analyst. Thus, data **"**is one of the most valuable assets of today's businesses and timely and accurate analysis of available data is essential for making the right decisions and competing in today's ever-changing real word environment. In general, there are three types of data analysis challenges, which include analytics, communication and application. Therefore, Automating data analysis refers to the task of discovering a relationship between a numbers of attributes and representing this relationship in the form of a model."

Data Mining refers to extracting, mining, finding, summarization and simplification the knowledge hidden from big database or defining the relationship hidden among the objects or attributes in a huge databases. Techniques of data mining vary according to the purpose of mining process. Usually, data mining tasks are divided into two categories: Description and Prediction. First, the description tasks include clustering, summarization, mining associations, and sequence discovery that are utilized to find understandable patterns of data (Jiaw *et al*., 2013).

Second, prediction is one of data mining techniques that have the ability to find unknown values of a target variable based on the values of some other variables. Furthermore, prediction is used to make future- plans using special techniques across different periods of time in extrapolating the development of assumptions about future conditions and opportunities. In short, it estimates activity in the future, taking into account all the possible factors that affect the activity. Prediction techniques in data mining are widely used to support optimizing of future decision-making in many different fields such as Marketing, Finance, Telecommunications, Healthcare and Medical Diagnosis. In this work, prediction is used to provide future information for business sectors.

This work not only presents and explores of the most eight widely used prediction techniques in the field of business sector. General properties and summary of each technique is introduced together with their advantages and disadvantages. Most importantly, the analysis depends up on the parameters that are being used for building a prediction model for each one and classifying them according to their main and secondary parameters. Furthermore, the presence and absence have also compared between them in order to better identify the shared parameters among the above techniques. Further, the main and optional steps of the prediction procedure are comparatively analyzed and produced in this paper.

The objective of this paper is to develop a smart, accurate and efficient prediction algorithm for huge and complex business databases by using cognitive agent, data mining and prediction techniques concepts, by taking the advantage of the cognitive agent technique, one can determine and select the most important features in order to reduce the time used in the predictor. Data mining techniques have the ability to deal with huge databases. Prediction techniques provide the ability to have a better way to look at future business behaviors and plans. This work combines Gradient Boosting Machine(GBM)and Develop Support Vector Machine(DSVM)based cognitive agent instead of decision trees, which GBM-based algorithms typically use, for prediction in huge business databases, it is determined to be more

accurate and effective predictor to achieve high accuracy, high speed of prediction (execution) and less cost.

The rest of the paper is structured as follows. Section 2 presents the related works Section 3 presents the predicate techniques used while in Section 4, the suggested tools used to building new predicator. Section 5 shows the challenges, steps of generated the proposed predicator (Agent-DKGBM). Section 6 shows experiments. Finally, the discussion and conclusion of the paper is presented in Section 7.

2. Related Works

This section briefly presents some of the recent research related to business sector using data mining and prediction techniques.

A study by Boyacioglu *et al*. (2009) planned to use several support vector machines, neural network techniques and multivariate statistical method in an attempt to investigate the impact of bank financial failure prediction problem in Turkey. The findings indicated that a multi-layer perception and learning vector quantization could be the most successful models in predicting the commercial failure of banks (Boya *et al*., 2009).

Lin *et al*.(2011) in their work combined both isometric feature mapping (ISOMAP) algorithm and support vector machines (SVM). This combination was used to predict the failure of companies founded on previous financial data. The results indicated that the ISOMAP has the best classification rate and the lowest incidence of Type II errors. In addition, there was a greater predictive accuracy that can be practically applied to identify possible financial crises early enough to potentially prevent them (Lin *et al*., 2011).

In another study conducted by Lu (2012) used Multivariate Adaptive Regression Splines (MARS), a nonlinear and non-parametric regression approaches to build sales predicting models for computer suppliers, to be able to make better sales management choices. The MARS prediction was able to develop useful information about the relationships between sales totals and the prediction variables also yields provided important prediction outcomes for making effective sales decisions and developing sales strategies (Lu *et al*., 2012).

Ticknor (2013) suggested using a Bayesian regularized artificial neural network to predict the changes and behaviors in the financial market. Microsoft Crop and Goldman Sachs Group Inc stock were used to assess efficiency of the proposed model. The findings showed that, the model has achieved more progressive prediction levels and it does not require preprocessing of the data, any cycle analysis or a seasonality testing (Tick., 2013).

He *et al*. (2014) used random sampling to improve SVM method and choose the F-measure to gauge the predictive power. The findings indicated that, the combination of the random sampling method and the SVM model significantly improved the predictive power that banks can use to predict the loss of customers more correctly (He *et al*., 2014).

Farquad *et al*. (2014) proposed a hybrid approach with three stages to extract the rules from SVM for Client Relationship Management (CRM) purposes. (i) SVM-RFE (SVM-recursive feature elimination) is used during the first stage to temper the feature dataset. (ii) During the second stage, the changed dataset is used to excerpt the SVM model and support vectors. (iii) Naïve Byues Tree (NBTree) can be used to produce Rules of Prediction. In the end, the researchers determined that this approach provided better results than the other techniques they tested (Farq *et al*., 2014).

3. Predicate Techniques

3.1 Analysis of Predicate Techniques

Prediction tries to solve a variety of problems by producing many technical works to find optimal or reasonable solutions for a specific problem. In this section, the major properties of eight prediction algorithms have been considered and a comparison among them is shown in Table 1.

A. Classification and Regression Tree (CART)

CART is one of the DTs techniques that are used to classify data easily in a more understandable form. To classify a data problem, the value of the target variable (Y) is found by using some interesting variable (X). It recursively splits the data from top to bottom to build the tree. Each branch represents a question about the value of one of the X variables to specify which direction the child nodes are to follow, right or left. If there are no more questions to ask in which specific direction to grow, it will terminate into a terminal node. It makes splits dependent only on one variable in each level (Romal, 2004).As a result, more accurate split from a combination of variables may be lost. If the number of variables is high, too many levels will be needed and more computational time will be required.

B. CHi-squared Automatic Interaction Detection (CHAID)

CHAID is another DTs techniques that allows a multi split of the parent node i.e., number of child nodes can be more than two for each parent node. It transforms continuous predictors to categorical ones (Kass.,1980). It allows multi-split, and therefore, it gives all variables more chances to appear in the analysis, regardless of their type, so it is useful with large dataset specially market segmentation. In this technique, nominal or ordinal categorical variables are only accepted; when variables are continuous they will be transformed into ordinal variables, which requires more preprocessing time, and therefore, a large number of categories can be expected from this translation with small amount of effect in the targeted predictor and makes additional merging steps before splitting the data set. In addition, it needs many user specific parameters such as, alpha-level merge (α alpha merge), alpha-level split-merge (α alpha split-merge) that causes a decrease in the automatic procedure.

C. Exchange Chi squared Automatic Interaction Detection (ECHAID)

Like CHAID, it allows multi split of the parent node, i.e. number of child nodes can be more than two for each parent node. Unlike CHAID, it merges more steps; comprehensively search procedure merges any similar pair until only single pair remains and compares p-*value* with the previous step rather than with user specific parameter (Stat.,2010). It needs less user specific parameters compared to CHAID, no alpha-level merge (α alpha merge) or alpha-level split-merge (α alpha split-merge) are needed providing more automatic operations.

D. Random Forest Regression & Classification (RFRC)

In a dataset, there are many variables representing possible permutations for nested split operations and it is possible to lose the optimal ones. It uses random selection, which is made by a random forest for subsets from variables to build a tree, followed by choosing another subset randomly to build another tree. Finally, a forest of trees resulting from the training set, each record is assigned to a class base on the higher voting principle. It is faster than bagging or boosting. In addition, it gives a useful internal estimate of error, strength, correlation, variable importance, it is simple and easily parallelized (Sut., 2011), but the required training time can be many hours or even days of computation, specifically for huge and complex data sets (Kurs., 2014).

E. Multivariate Adaptive Regression Splines (MARS)

MARS is a data driven regression procedure that build a model based on “divide and conquer" principle from a number of basis function equations and coefficients, each equation pertaining to a region in the input space. This approach can handle multi-dimensional data as variables that represents the main problem in other techniques (Isla *et al*., 2015). It does not have a tree like CART or CHAID techniques, a series of equations, which perform regression tasks; thus, it depends totally on mathematical functions in finding the optimum solution. It has a fast prediction results with new unseen data, but sometimes, it suffers from discontinuity in sub region boundaries that may affect the accuracy (Mans., 2014).

F. Boosted Tree Classifiers and Regression (BTCR)

BTCR is a finite loop of CART's operations that produce a sequence of binary trees with forward strategy. Like RFRC, it builds a number of trees to reach optimal accuracy by covering all possible variable choices, but it does not depend on bootstrap sampling or selection variable by randomness. Each iteration tree is built to predict a residual of the previous tree. It prevents the tree from growing without control. In each epoch, the size of the current tree is detected to a fixed value (Stat., 2010) (Mans.,2014) (Eli *et al*., 2008) Furthermore, it avoids overfitting by limiting the number of iterations (Mans.,2014).The limited and incorrect size of a tree affects BTCR prediction performance because it does not handle interaction between variables. In addition, poor prediction can happen with a small number of samples. There is also no need for data transformation or elimination of outliers and it can automatically handle interaction effects between predictors

G.Bayesian Neural Networks Classifier (BNNC)

BNNC is one of the classification methods that uses both the probability theory and graph theory to solve the problem of non-deterministic relation between variables set X and target class Y. Graph is used as a qualitative part, the node representing the variables while the arcs representing the dependency between them. Arcs must be directed and without any cycle. A set of parameters is used as a quantitative part to represent the conditional probability distributions (Gorul.,2011).Despite the consumption time in constructing the network, which requires a large amount of time and efforts, it is more flexible than the naıve Bayes classifiers by allowing some dependency between variables. It can deal with incomplete data and because the data is combined with prior knowledge, the method is quite robust to model overfitting (Gorul., 2011).

H.Support Vector Machine (SVM)

SVM is a classification approach that relies upon the decision boundary principle to determine the class value depending on the location of the point that characterizes the variables. SVM mitigates the binary class problem by applying linear separable input space. For non-linear separable input space, the variables are transformed into another linear space. This concept allows locate the optimal line of decision with the maximum margin to resolve the issue of overfitting, which can be overburdening(Gorul,2011; Jin.,2014).It requires preprocessing time for the variables' transformation from non-linear space to linear space;also,the user must provide parameters including the type of kernel function and the cost function C.

TABLE 1: Comparison among Main Techniques of Predication

Techniques	Advantages	Disadvantages	primary parameters	Secondary Parameters
CART	-Automatic class balancing. Automatic missing Values handling - Easily handle outliers –with justifications	- Don't have DTs - Splits only by one variable. at a time.	Y , X , *w* , *f*	- Maximum Tree Depth - Minimum Node Size
CHAID	- Allows multiple split. Handles market segmentation	-Variables transformation - Uses three user specific parameters. -Time for merge steps	Y , X, *W* , *F* , Alpha_Merge, Alpha_Split-Merge, Alpha_Split	- Maximum Tree Depth - Minimum Node Size
ECHAID	-Allows multi split. - handle market segmentation -Needs less user specific parameters than CHAID	-Variables transformation - Tow user specific parameters . -Time for merge steps	Y , X , *w* , *f* , alpha_split	- Maximum Tree Depth - Minimum Node Size
RFRC	- Good accuracy - Robust to outliers and noise. - Faster than bagging or boosting. -Simple and easily parallelized.	Requires time from hours to even days of computation, especially for larger s datasets.	Y , X , *mtry* , *ntree*	- Maximum Tree Depth - Minimum Node Size
MARS	- No user specific parameters. - No variable transformation - More flexible – Automatic variable selection – Fast prediction. but error pone	-Discontinuity in sub-region boundaries that effect on accuracy. -Needs backward steps to fix overfitting	Y , X	*M*
BTCR	- Limit number of iterations . - Handle many type of target variable y . - No need for data transformation or elimination of outliers.	- Poor prediction with incorrect tree size limit and with little number of samples	Y , X , *J* , *B* , *nmin*	-
BNNC	-More flexible than nauve bayes classifiers - Well suited to dealing with incomplete data - Quite robust to model overfitting	-Time consuming for constructing the network - Requires a large amount of time and efforts .	Y , X , P(X), P(Y)	-
SVM	-Global minimum objective function. - Maximizing the margin to control capacity -Handles categorical data with dummy	-Variables transformation - Uses two user specific parameters.	Y , X , *K* , *C*	-

Where Y is the targeted variable, X is the interested variables, w is the case weight, f is the frequency weigh, mtry is the randomly sample of the predictors, ntree is the number of trees, B is the number of iterations , nmin is the minimum number of samples in a node , M is the number of Mars terms , P(X) is the prior probability of X, P(Y) is the prior probability of Y, K, C are the kernel functions.

3.2 Analysis of the Main Parameters of Each Prediction Technique

Each one of the prediction techniques has a number of parameters that can be primary factors in building the prediction model or they can be secondary ones, which contribute to the primary parameters to provide an optimum solution. To build a new prediction technique, it is necessary to analyze these parameters and make the comparison for the presence and absence of each one to determine their effects in the prediction techniques. Table 2 shows compression between prediction techniques parameters.

TABLE 2: Comparison among the Main Parameters, which affect Prediction Techniques

Parameters	***Y***	***X***	***w***	***f***	**alpha_merge**	**alpha_split-merge**	**alpha_split**	***mtry***	***ntree***	***J***	***P(X)***	***P(X)***	***K***	***C***	**Max Tree Depth**	**Min Node Size**	**Mars Terms**
CART	✓	✓	✓	✓	×	×	×	×	×	×	×	×	×	×	✓	✓	×
CHAID	✓	✓	✓	✓	✓	✓	✓	×	×	×	×	×	×	×	✓	✓	×
ECHAID	✓	✓	✓	✓	×	×	✓	×	×	×	×	×	×	×	✓	✓	×
RFRC	✓	✓	✓	✓	×	×	×	✓	✓	×	×	×	×	×	✓	✓	×
MARS	✓	✓	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓
BTCR	✓	✓	✓	×	×	×	×	×	✓	✓	×	×	×	×	×	✓	×
BNNC	✓	✓	×	×	×	×	×	×	×	×	✓	✓	×	×	×	×	×
SVM	✓	✓	×	×	×	×	×	×	×	×	×	×	✓	✓	×	×	×

3.3 The Main Prediction Techniques Steps

To build a new prediction technique with standard procedures, it is also necessary to consider and analyze all the steps of the current prediction techniques to choose the appropriate steps from all these methods that can help building and developing of a new prediction method, the steps of the prediction techniques are given in Table 3.

TABLE 3: Comparison among the Steps of the Main Prediction Techniques

Steps	Specifies X , Y	Variable transformation	Merges categories	Chooses random variable subset	Finds split variable	Finds split condition	Splits parent node	Check s splitting rule	Pruning	Comparing Trees	Caculates prior pro. P(X)P(Y)	Finds optimal hyperline splitter
CART	✓	×	×	×	✓	✓	✓	✓	✓	×	×	×
CHAID	✓	✓	✓	×	✓	✓	✓	✓	✓	×	×	×
ECHAID	✓	✓	✓	×	✓	✓	✓	✓	✓	×	×	×
RFRC	✓	×	×	✓	✓	✓	✓	✓	×	✓	×	×
MARS	✓	×	×	×	✓	×	×	×	✓	×	×	×
BTCR	✓	×	×	×	✓	✓	✓	✓	✓	✓	×	×
BNNC	✓	×	×	×	×	×	×	×	×	×	✓	×
SVM	✓	✓	×	×	×	×	×	×	×	×	×	✓

Where Y is the target variable; X is the interest variables; P(X) is the prior probability of X, and P(Y) is the prior probability of Y.

From experiments analysis of these prediction techniques, in Table 3, it can be observed that the techniques that are not dependent on randomization such as, BTCR provided better results compared to the techniques using mathematical basis such as, MARS, which is more powerful and faster. MARS provides an optimal solution because it utilizes features of mathematics such as, linear combination, simplification, derivatives and integration. In addition, from the analysis it can be determined that similar parameters are shared among all the prediction techniques such as, the target variable Y, the interest variable X and specified them as a higher priority. In this work, the proposed Agent-DKGBM algorithm was suggested, which is utilized features of mathematics such as, *Hyperbolic functions*, *Polynomial functions* and *Gaussian mixture.* The proposed algorithm will be compared with some of the previous predict techniques and is presented in the experiments results section. This work will compare with all techniques expect the techniques based of the probability (CHAID, E CHAID and BNNC).

4. Main tools for Novel Predictor

4.1 Agents

Agents can be defined as software that can interact within environments through sensors and actuators as shown in Figure 1. The agent has many properties:

- Reactive (i.e., Responds in a timely fashion to changes in the environment),
- Autonomous (i.e., Exercises control over its own actions),
- Goal-oriented (i.e., Does not simply act in response to the environment),
- Temporally continuous (i.e., a continuously running process),
- Communicative (i.e., Communicates with other agents, perhaps including people),
- Learning or adaptive (i.e., Changes behavior based on previous experience),
- Mobile (i.e., the ability to move in space),
- Flexible : Actions are not given
- Character: Credible 'personality' and emotional state ."

In this work, the agent has the capability to handle the steps of Business Intelligence and Analytics automatically. The prime role of the agent is to fetch databases from the environment as a sensory input, pass the different stages of Meta Knowledge predictor (MKP), take decisions when necessary, and then stores the results in a knowledge base file as an output. This automation process enables non-experts to easily use the system and facilitate the work for experts as well.

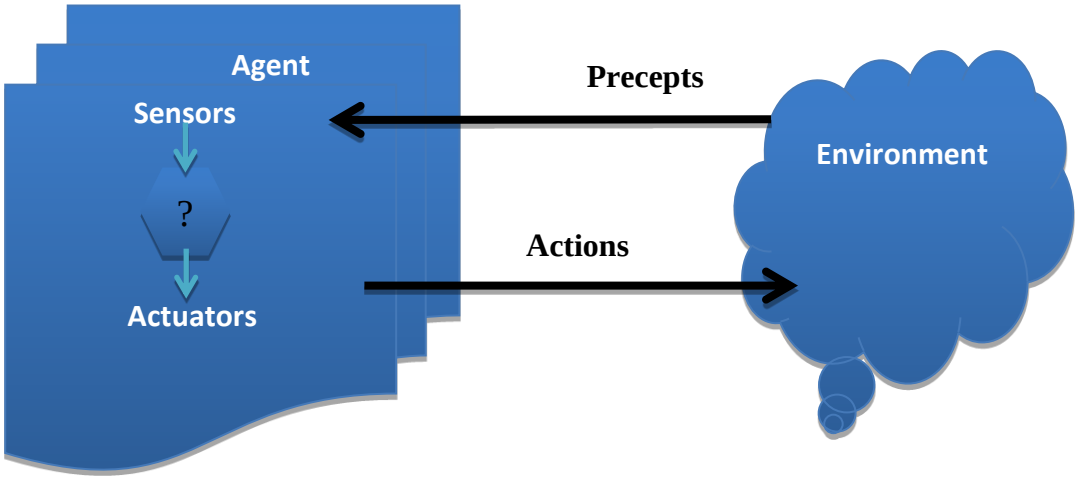


Figure 1. The General Structure of Agents.

In general, there are two types of agents, Reactive and Cognitive agents. First, the *Reactive agent* is very limited in what they can do, also they do not have the capability to plan and coordinate between themselves or set and understand specific goals; they simply respond to events when they occur. This does not prevent them from having a role to play in producing intelligent behavior. The reactive school of thought is that it is not necessary for agents to be individually intelligent. However, they can work together to solve complex problems. Second, The *Cognitive Agents* seeks to build agents that exhibit intelligence in some manner. In this method, individual agents have goals, and can develop plans on how to achieve them. They use more sophisticated communication mechanisms and can intentionally coordinate their activities. They also map their environment in some manner using an internal representation or knowledge base that they can refer and update through learning mechanisms in order to assist guide their decisions and actions. As a result, the Cognitive agents are more flexible in their behaviors compared to Reactive agents and is used in this work. A compression between the two types of agents is summarized in Table 4.

Table 4: Reactive Agents versus Cognitive Agents (Sheh., 2014)

Reactive Agents	Cognitive Agents
Use simple behaviors	Use complex behaviors
Have low complexity	Have high complexity
Are not capable of foreseeing the future	Anticipate what is going to be happened in the future
Do not have goals	Have specific goals
not capable to plan or coordinate amongst themselves	capable to plan and coordinate with each other
Have no representation of the environment	The ability to map their environment
Do not adapt or learn	Exhibit learned behavior
Can work together to resolve complex problems	Can resolve complex problems both working together and/or individually

4.2. Gradient Boosting Machine (GBM)

GBM is a powerful and brilliant prediction technique which utilize boosting concept for reducing error in prediction. GBM algorithm builds an additive model for minimizing residual which represent the value of subtracting each target value and mean of target (Max., 2013). The target value for each data record is re_estimated at each iteration, the aim of that to give the new tree its contribution (Nate.,2013). It provides a set of weak prediction models. These models typically utilize decision trees built upon a stepwise fashion and it streamlines the process to optimize arrangements of an arbitrary differentiable function.

The linear combination of Binary Regression Trees (BRT) represents the final model of GBM. The performance of building process is the best, if it proceeds slowly and reducing the contribution of each BRT by the learning rate, which has value of less than one. The results of GBM usually are much more stable and accurate than a single BRT model (Nate.,2013). It uses BRT for learning rate and it requires four parameters that have the major effect on its performance and behavior (Max., 2013) (Nate., 2013) (Fros *et al*., 2015). These parameters include maximum number of trees, learning rate, when the value is less than 0.1, it is called *Shrinkage* (Shrinkage < 0.1) (Fros *et al*., 2015), maximum number of terminal nodes in the BRT, and minimum number of data records in a terminal node.

5. The Proposed Agent-DKGBM Predictor

The main goal of this study is to construct a new Agent-DKGBM algorithm that can achieve less cost, the high speed of prediction (execution time) and high accuracy, which are the three main challenges in prediction, as shown in Figure 2. This will not only allow the business sectors to deal with very huge databases and the ability to plan for future, it would help to better understand the needs of their customers better, to predict their wants, demands and maximize their profits..

The Agent-DKGBM algorithm executes in two phases. In the first phase, building the cognitive agent, the primary goal is to prepare the database for the second phase - searching the business databases. During the first phase, the agent selects one of the business databases from the bank of business databases, choose the suitable type (i.e., Hyperbolic functions, Polynomial functions, Gaussian mixture) as a kernel of Develop Support Vector Regression (DSVR), and determine the optimal parameters of DSVR and DKGBM.

The second phase consists of three stages, which include splitting the business databases into training and testing datasets by using 10-fold cross validation. In the

second stage, a DKGBM model using the training data set is built to replace the kernel of Gradient Boosting Machine (GBM), typically using Decision Trees (DTs) by DSVM in order to increase the accuracy, reduce the execution time and cost in the DKGBM model. Finally, the results of the DKGBM would be verified based on the testing data set.

Figure 2. Relationships among the Main Three Challenges

5.1 This Study Embarks On the Following Objectives

- Analyze the parameters of eight data mining prediction techniques and determine each of the parameters consider effective decision-making. Then establishing a new prediction algorithm called Agent-DKGBM based on the previously analyzed approaches and building the fundamental modeling to prove the theory and establish its viability.
- Prove that can satisfy the three main challenges of prediction (the high accuracy, the high speed of prediction (execution) and less cost
- Assess the algorithm and framework analysis of the proposed strategy.
- Make recommendations for both sides: The first for business/marketing about predict of the close price of the marketing for the next days to take all the necessary requirements, and the another report show the optimal parameters for DSVM and DKGBM generated by the Agent.
- Patent the resulting algorithms if Agent-DKGBM hit rates exceed those of current.

5.2 Research Methodology

The following points represent the main steps of research methodology, Algorithm 1 summarizes the main steps used to generate the new Agent-DKGBM predictor; Figure 3 presents the architecture of Agent-DKGBM algorithm, while the research works, activities are shown in Figure 4.

A. Determine the main parameters in each prediction techniques that effect in the prediction

B. Develop the GBM by replacing the kernel of GBM in DTs by using DSVM to reduce the execution time and increase the accuracy.

C. Building a cognitive agent that prepares the database of the second phase through a search of the business environment, the agent performs multi tasks in parallel:

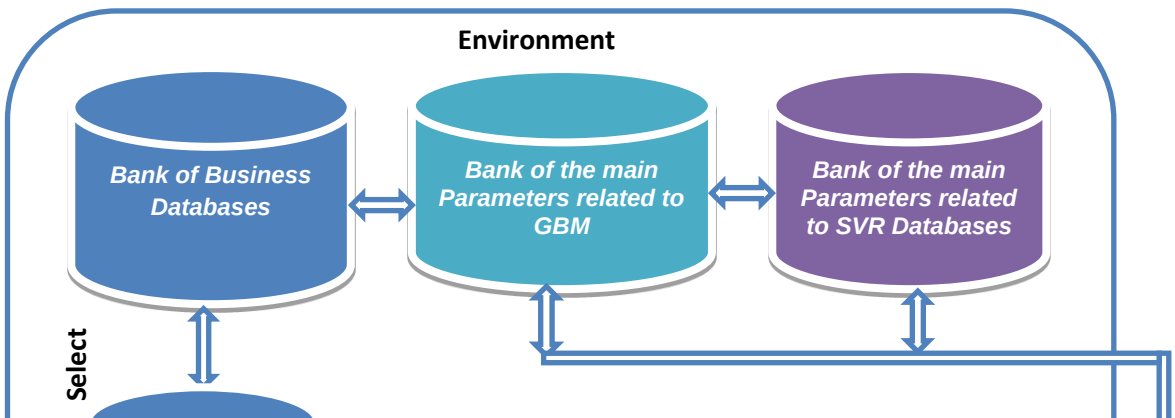
(i) Select one of the business databases.

(ii) Choose the suitable kernel type (i.e., Hyperbolic functions, Polynomial functions, Gaussian mixture) as a kernel for Develop Support Vector Regression (DSVR).

(iii) Determine the optimal parameters of DSVR and DKGBM

D. Design a new Agent-DKGBM algorithm summarized in Algorithm 1.

E. Compare the result of the Agent-DKGBM with all the other prediction algorithms.


Environment
Bank of Business Databases
Bank of the main Parameters related to GBM
Bank of the main Parameters related to SVR Databases
Select

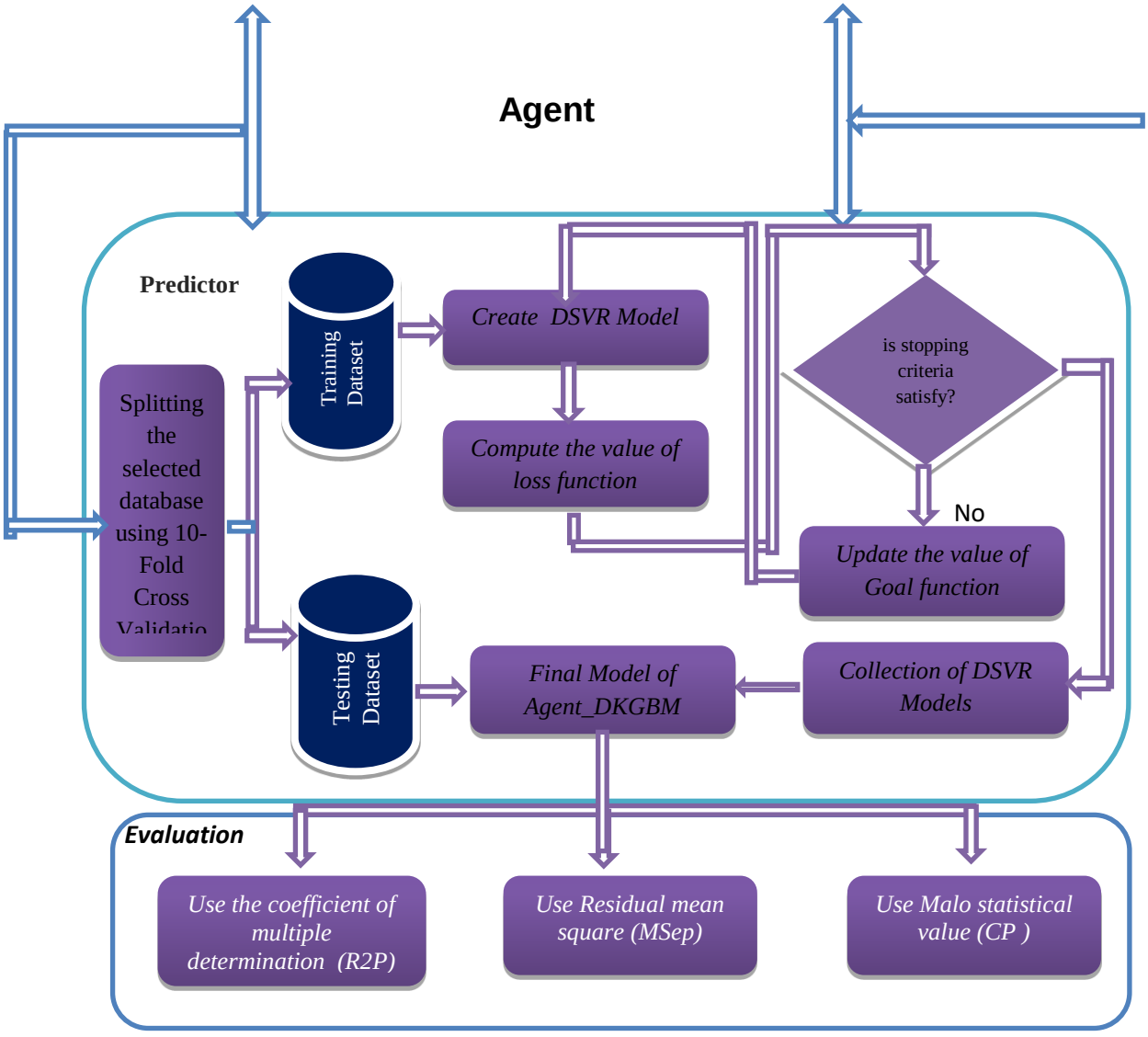


Figure.3. The architecture of Agent-DKGBM

Algorithm 1: Agent-DKGBM

Input: *Collection of Databases related to many banks.*

Output: *Optimal predicted values*

Method:

• ***Step 1:*** *Build the Cognitive agent to search and determine the main perimeters the business environment*

- *Select one of the business databases.*
- *Choose the suitable kernel type (hyperbolic function, Tansh Function, Gaussian Mixture) as kernel of Develop Support Vector Regression (DSVR)*

Algorithm 2 : DSVR

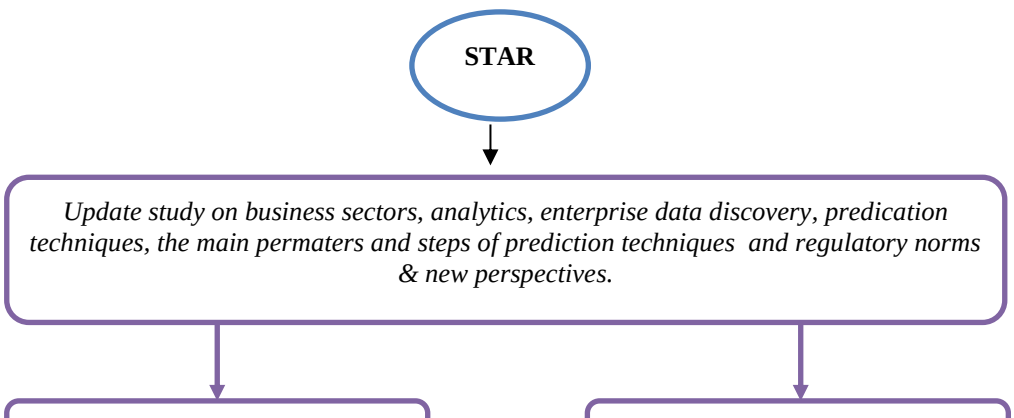
Input: *Training Dataset.*

Output: *Collection of final DSVR Models*

Method:

- *Based on the Kernel function, the kernel's parameters, and the soft margin parameter Swere selected by Agent in algorithm 1.*
- *Based on the group of parameters that were set in the algorithm 1.*

- ***Step 1:*** *if the number of DSVR_MAX less than Max number of iteration*
- *Test Counter of records (Cr) less than or equal max number of the training records.*
- *For each training pair (X_i, G_k) // $g_{1,\dots,k}$: values of Goal*
  - *Compute the prediction function as sum of cross products with the new sample*


STAR
Update study on business sectors, analytics, enterprise data discovery, predication techniques, the main permaters and steps of prediction techniques and regulatory norms & new perspectives.

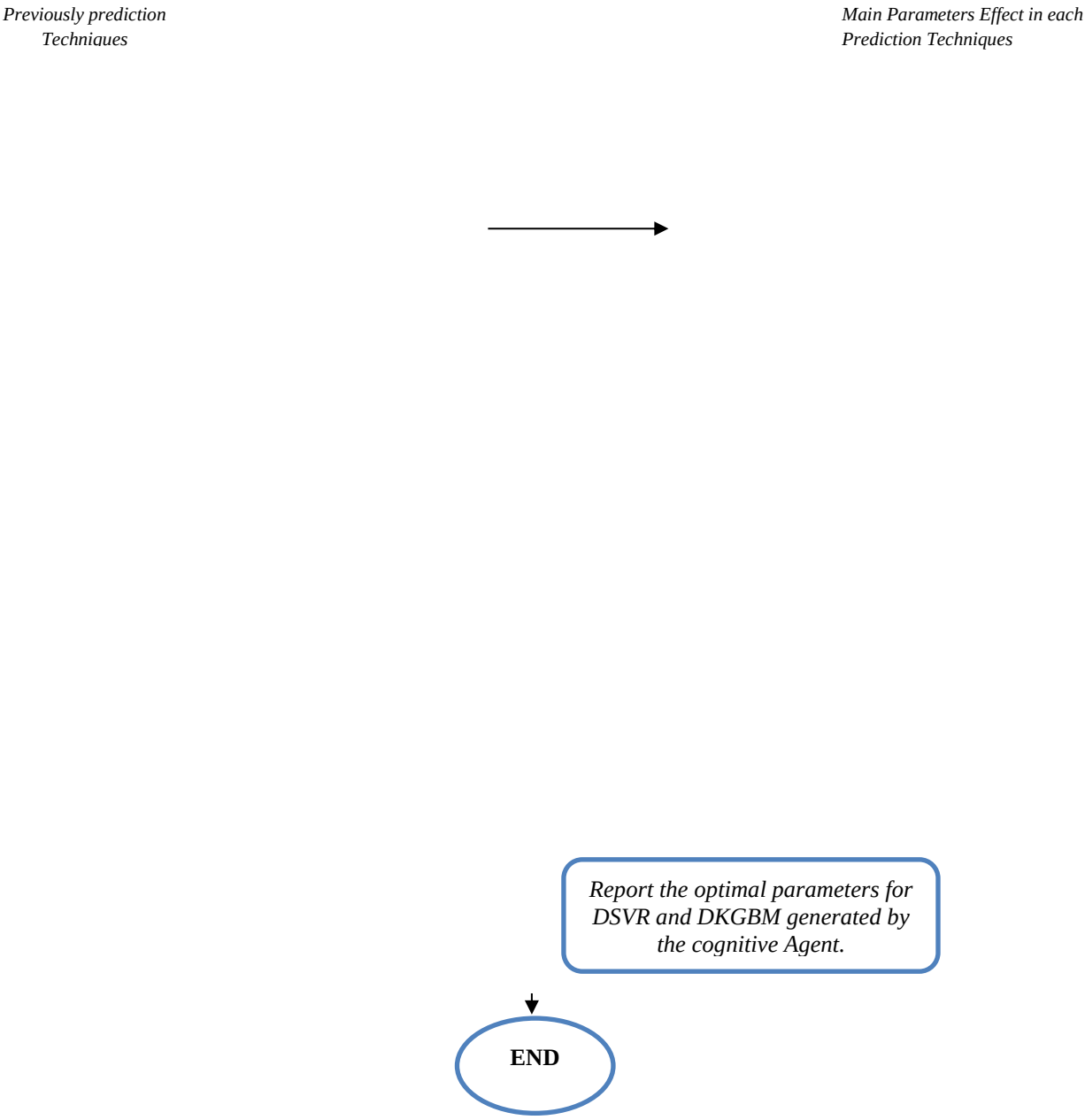


Figure 4: Flow Chart of the Research Work Activities

6. Experimental Results

This section aims to describe the implantations phases were used in building the Agent-DKGBM predictor and compare its performance with other predictors.

6.1 Performance Metrics

This work evaluated the performance of the Agent-DKGBM predictor. Al Rajhi Capital[1] that associated with Saudi Arabia market was selected as a real dynamic business datasets. This market deals with six standard features, which include *Open Price*, *High Price*, *Low Price*, *Volume*, *Close price and Turnover*, the close price is the main feature selected to be predicted, The results of Agent-DKGBM would be tested based on three error measures are used. Three measures, which include the coefficient of multiple determinations (R^2P), Residual mean square (MScp) and Malo statistical value were calculated to analyze the differences between the actual and expected results. In addition, 10-fold cross validation is used to test and validate the results of the proposed predictor. Equations (1), (2) and (3), were used for calculating these metrics as given below:

$$R^2P = \frac{SSR(X1,X2,\ldots..,Xp)}{SST} = 1 - \frac{SSe(X1,X2,\ldots..,Xp)}{SST} \quad (1)$$

$$MScp = \frac{SSe(X1,X2,\ldots..,Xp)}{n-p} \quad (2)$$

$$CP = \frac{SSe(X1,X2,\ldots..,Xp)}{MSR(X1,X2,\ldots..,Xm)} - (n-2p) \quad (3)$$

Where, *SSR* is measure of explained variation, *SSe* is measure of unexplained variation, *SST* is the measure of total variation (SSE+SSR), *MSE* is the mean square error, *n* is the total number of test data , *p* is the correct number of predicted values by the Agent-DKBGM . Table 5 shows the optimal values of the parameters for the selected datasets, associated with the error measures.

Based on the three error measures were used in this work, 60% of the dataset for training and the rest for testing the model are the best splitting option for the dataset as shown in Table 5.

Table 5: Testing Agent-DKGBM based on the 10-fold Cross Validation and Three Error Measures

[1] Al Rajhi Capital : http://www.alrajhi-capital.com/en/Pages/default.aspx

Predictor	Al- Rajhi Capital Dataset		R^2P	MScp	CP
	Training	Testing			
Agent_DKGBM	90%	10%	0.324	0.163	0.873
	80%	20%	0.835	0.414	0.882
	70%	30%	0.211	0.311	0.724
	60%	**40%**	**0.9457**	**0.016**	**0.097**
	50%	50%	0.493	0.562	0.626
	40%	60%	0.845	0.822	0.995
	30%	70%	0.548	0.373	0.142
	20%	80%	0.759	0.423	0.724
	10%	90%	0.893	0.763	0.485

The results in Table 6 indicates that the Gaussian mixture (d-dimensions) is the best kernel function in both time and value of error and has d-dimensions as explained in the final paragraph of step number two in Algorithm 2. Due to this, the agent selects the Gaussian mixture as a kernel function of DSVR. Therefore, the Gaussian mixture can be chosen as a kernel for DSVR that used as based in DKGBM rather than DT. Table 7 presents a comparison between the actual and predict values of the close price by Agent-DKGBM for 800 samples.

Table 6: Testing Agent-DKGBM based on the Different Kernel

Table 6: Testing Agent-DKGBM based on the Different Kernel

Predictor	Kernel Function	# of Iterations	Time	Value of Loss Function
Agent_DKGBM	***Tanh function***	68	36	0.825
	Triple Function	30	112	0.726
	Logistic Function	150	125	0.930
	Gaussian mixture	25	22	0.081

Table 7: Compare between the actual and predict values of the close price by Agent-DKGBM

# of observations	Actual(Close)	Agent_DKGBM (Close)	Residuals

1	70.160	69.671	0.489
2	70.160	70.115	0.045
3	69.690	70.115	-0.425
4	69.690	69.653	0.037
5	70.390	69.653	0.737
6	70.620	70.341	0.279
7	70.620	70.567	0.053
8	70.390	70.567	-0.177
9	71.080	70.341	0.739
10	71.310	71.020	0.290
.....	.....	.....	.....
.....	.....	.....	.....
796	66.930	68.522	-1.592
797	67.390	66.939	0.451
798	67.390	67.391	-0.001
799	67.390	67.391	-0.001
800	67.390	67.391	-0.001

In Figure 5, it can be observed that the predicted output follow the actual output showing that the developed Agent- DKGBM algorithm is capable of successfully predicting future values. While, Figure 6 shows the histogram residual of the closed value.

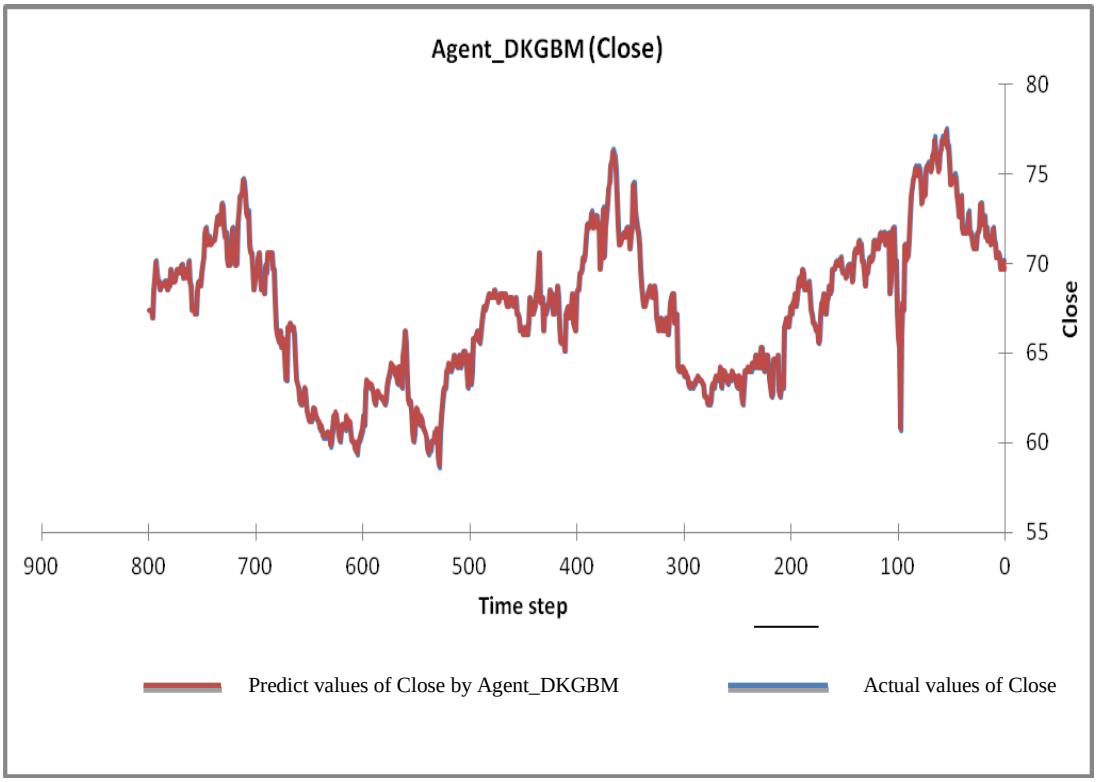


Figure 5: The actual and predicted Values of Close generated by Agent_DKGBM

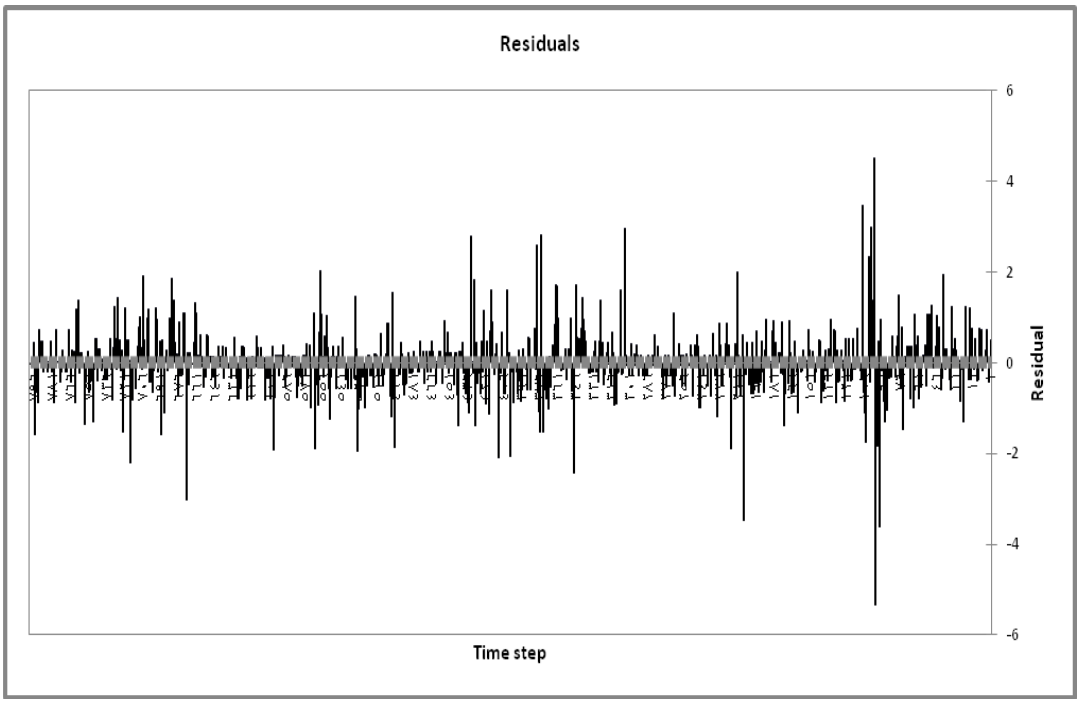


Figure 6: Residual of Close values

One of the main factors in any business sector is the turnover and any changes of these values is very important in determining the weight or rank of that market compared to other ones internationally, and therefore, the histogram in Figure 7 shows the relation between the actual and predicated values of turnover which generated by the Agent-DKGBM..

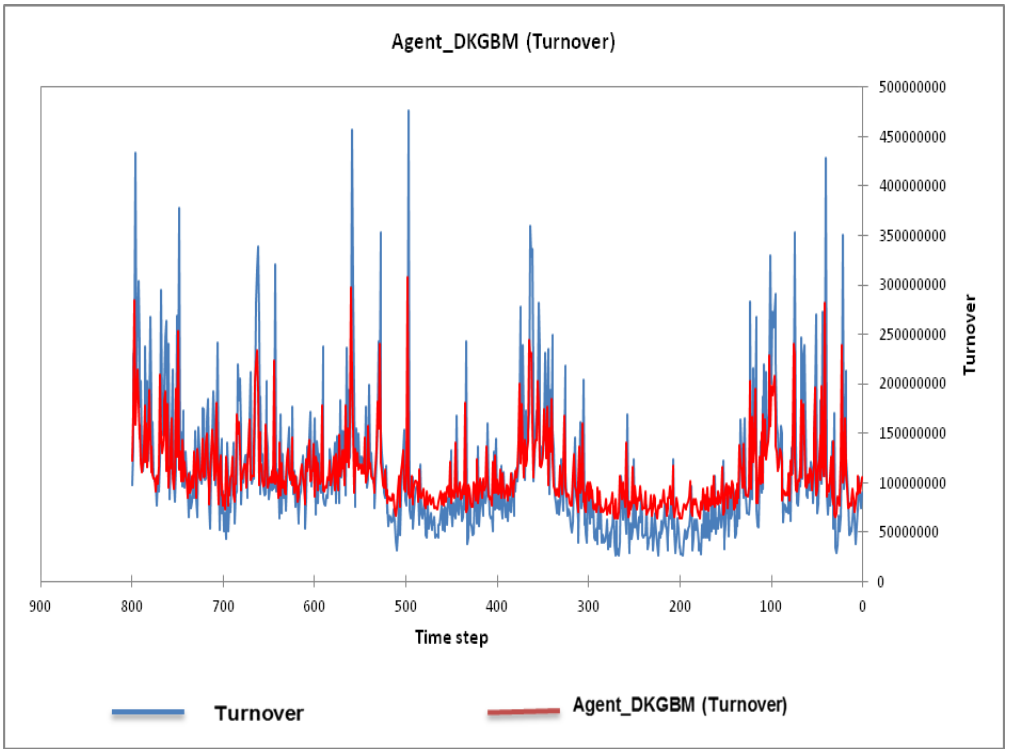


Figure 7: The actual and predicted Values of Turnover generated by Agent_DKGBM

6.2 Compressions between Prediction Techniques and the Agent_DKGBM

To evaluate the performance of the proposed algorithm, experiments conducted to determine the relative significant of prediction techniques on the error measures as shown in Table 8.

A. CART

The main parameters of that predictor are, the total number of trees is143, Max depth of each tree is 13, the number of optimal tree is71, the actual depth is10 and the number of terminal nodes in the optimal tree is82. The fitted model of that predictor appears by green color as shown in Figure 9.

B. SVM

The main parameters of that predicator are, Epsilon = 0.001, C = 0.33274212, Gamma = 0.00332742, Coef0 = 35.9381366, P = 0.00010197, the number of point evaluated during search is1127, the minimum error is 0.168165. Fitted model of that predictor appears by red color as shown in Figure 9.

C. MARS

The main parameters of that predicator are, P1=59.3583, P2=0.991792, P3=0.209281, P3= - 0.105715, while the basis functions X1=max (0.High-59.75), X2= max (0.High-73.5), and X3= max (0.High-71.5), final model Y= 59.3583+0.991792*X1+0.209281*X2- 0.105715*X3. The fitted model of that predictor appears by pink color as shown in Figure 9.

D. RFCR

The main parameters of that predicator are, the Max N of Trees is 200, Target is Close, the number of predictor's is5, the maximum depth of any tree is 14, and the minimum nonterminal node size is 2. In general, the results of RFCR are very close to SVM. Fitted model of that predictor appears by red color as shown in Figure 9.

E. BTCR or Tree Net

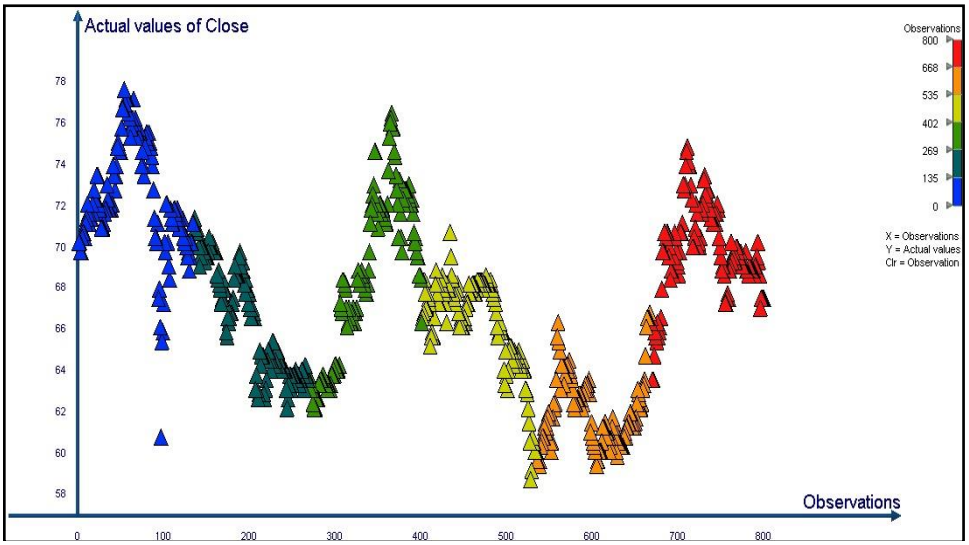
the main parameters of that predictor are, the Max N of Trees is 400, Target= Close, the number of predictors is 5, the minimum error with the training data occurs with 400 trees, the minimum error with the test data occurs with 11 trees, the maximum depth of any tree is 10. In general Tree Net and GBM behaviours are very close to each other from both the structure and results sides. Fitted model of that predictor appears by blue color as shown in Figure 9.

F. GBM

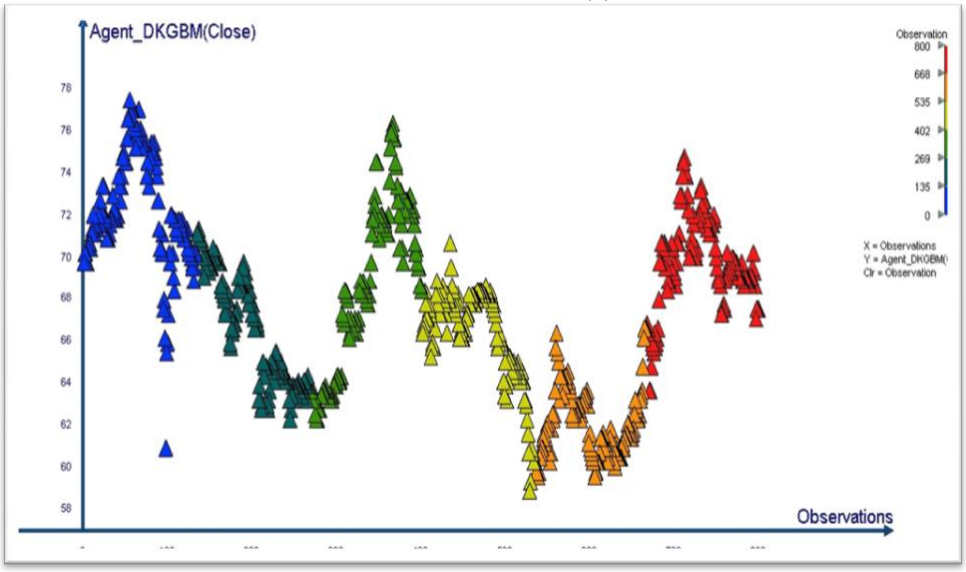
The main parameters of that predictor are, the max N of Trees is 400, Target= Close, number of predictors is5, the maximum number of terminal nodes in every single binary tree that controls the number of rules is 16, .the minimum number of samples in each terminal node that effect on the coverage of the rule of that node is 10, the learning rate (Shrinkage) =0.612. In general, the fitted model of that predictor appears by blue color in Figure 9.

G. Agent-DKGBM

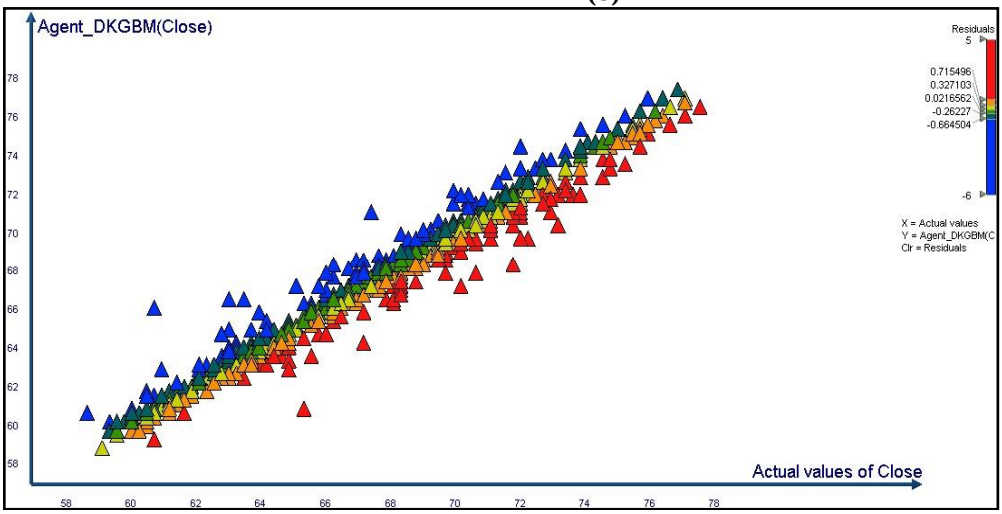
The main parameters of that predictor are, the Max of iterations is 200, Goal = Predict the values of Close price, the number of predictors is5, the rate of training dataset is 60%, the rate of testing dataset is 40%, activation function of DSVR=Gaussian mixture, learning rate (Shrinkage) =0.011, base function of GBM=DSVR, the Max number of iteration in SVR is25. Figure 8 shows, the actual and predicted values generated by the Agent-DKGBM for all the samples were used in the training and testing phases. The fitted model of that predictor appears by discreet black color as shown in Figure 9.



(a)



(b)



(c)

Figure 8. The Results of Agent-DKGBM for Al Rajhi Market that included. (a) Relationship between the actual values of close price and observations, (b) The relationship between the predicted values of close price and observations, (c) The

relationship among the actual values of close price, the predicted values of close price and residuals.

Table 8: Compare the Accuracy among the Agent_DKGBM and other Predictors

Predicator	R^2P	MScp	Cp
CART	0.561	0.693	0.286
SVM	0.857	0.526	0.553
MARS	0.766	0.497	0.423
RFCR	0.853	0.526	0.562
Tree Net	0.870	0.792	0.301
GBM	0.891	0.787	0.296
Agent_DKGBM	**0.945**	**0.0167**	***0.*097**

As can be seen in Figure 9, the proposed Agent-DKGBM algorithm is capable of both successfully and accurately forecasting future values. The proposed algorithm will also allow the business sector to have a better way to look at future business behaviors and their future decisions making compare to other prediction techniques.

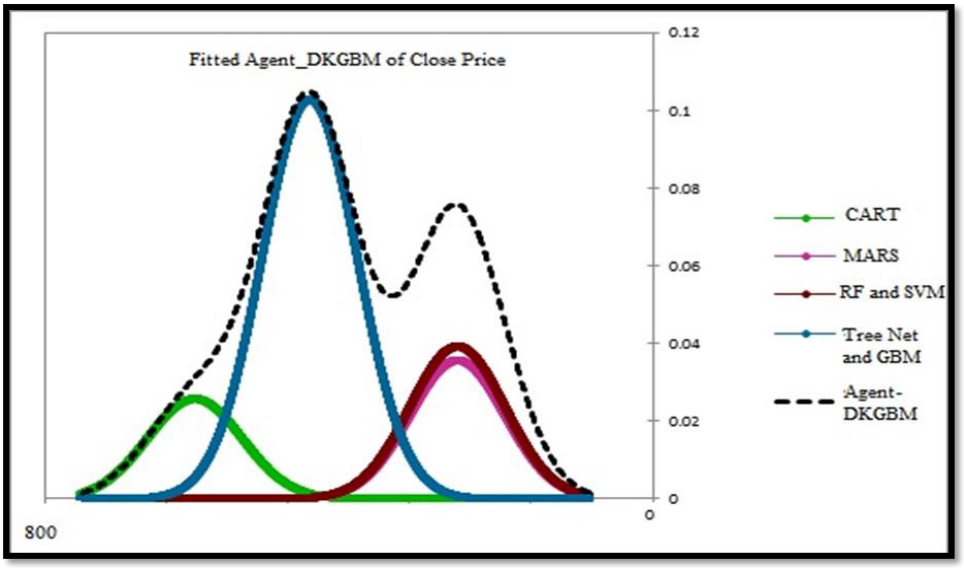
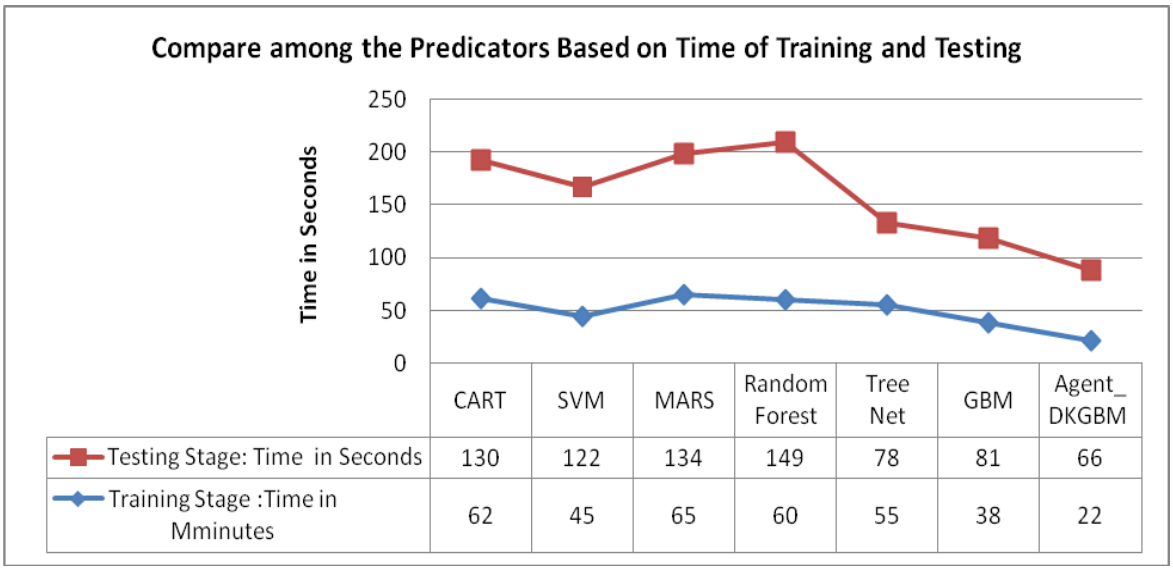


Figure 9. Comparison among the Fitted Models of Closed price Values

The results in Table 9 indicate that the Agent-DKGBM is capable of predicting with less execution time than other prediction techniques. The required time of 22 minutes in the training phase and the time of 66 seconds in the testing phase are both less than other techniques as shown in Table 9. A comparison among the predictors based on time of training and testing phases is shown in Figure 10.

Table 9: Compare the time among the Agent_DKGBM and other Predictors

Testing Stage: Time in Seconds	Training Stage :Time in Mminutes	Predicator
130	62	**CART**
122	45	**SVM**
83	55	**MARS**
189	60	**RFCR**
231	65	**Tree Net**
81	38	**GBM**
66	***22***	**Agent_DKGBM**



	CART	SVM	MARS	Random Forest	Tree Net	GBM	Agent_ DKGBM
Testing Stage: Time in Seconds	130	122	134	149	78	81	66
Training Stage :Time in Mminutes	62	45	65	60	55	38	22

Figure 10. Comparison among the predictors based on time of training and testing phases.

7. Conclusion and future works

This paper implemented Agent-DKGBM for predicting the closed price in business domain. In addition to that, it analyzed and compared some of the existing prediction techniques in an attempt to determine the main parameters that have the most important effects on their predictor. From the analysis, we found the techniques that are not dependent on randomization provided better results such as BTCR, while the ones using mathematical basis offered more powerful and faster solutions such as MARS. In the light of this, mathematical basis is used in the proposed Agent-DKGBM technique.

The Agent-DKGBM combine among the advantages of ***cognitive agent*** [i.e., it is used as pre-processing phase to prepare the database for the second phase, searching the business databases. During this phase, the cognitive agent selects one of the business databases, choosing the most suitable type i.e., Hyperbolic functions, Polynomial functions and Gaussian mixture) as a kernel of DSVR and determines the optimum parameters of the DSVR and DKGBM. The results of this stage show that Gaussian mixture is the best activation function for DSVR,, the best value of learning rate (Shrinkage) is 0.011, the based function of GBM is DSVR, the Max number of iteration in SVR is 25], ***10-fold cross validation*** (i.e., find the best split of database is 60% used in training stage, the rest used in testing the model), ***Developed Gradient Boosting Machine*** kernel through replacing Decision Trees (DTs) by DSVR (i.e., increase the accuracy and reduce the execution time) in the DKGBM model.

The results show that the Agent-DKGBM performs better than other prediction techniques in prediction the closed price in business domain. The coefficient of multiple determinations (R2P), Residual mean square (MScp) and Malo statistical value measures makes it obvious that the Agent-DKGBM performs better than other techniques. The results from Agent-DKGBM produce smaller estimation error compared to other prediction techniques. Thus, it can be concluded that Agent-DKGBM can allow the business sector to have a better way to look at future business behaviours and can play a vital role in their future decisions making. The results show that the proposed algorithm achieves an improvement in accuracy, speed of prediction and less cost. Therefore, the proposed Agent-DKGBM is promising choice compared to other prediction techniques. The experimental results also show that the proposed algorithm employed in this work overcomes some of the shortcomings in other

prediction techniques and it can be used in many different domains such as Finance, Marketing, Telecommunications and Medical Diagnosis.

The results also show that some of predictors give very close results to each other such as (SVM and RFCR) while some of them are similar in both the work structure and the results such as (GBM and Tree Net). As future work, we planning to develop Tree Net Algorithm by use optimization trees (i.e., Genetic Programming) as based function on it instead of decision trees. Using one of optimization algorithms such as swarm optimization, Ant Colony Optimization (ACO) and Genetic Algorithm (GA) to determine and select the most important features in order to reduce the time used in the predictor. The results would be evaluated for the new predictor based on another error measures such as MSR. RMSR, ARMS and Log RMSE or the measures based on confusion matrix, namely: "Accuracy (AC), recall or true positive rate (TP), precision (P), F-measure (considers both precision and recall) and Fb".

REFERENCE

Mans., 2014 Mansiaux, Y., & Carrat, F. (2014). Detection of independent associations in a large epidemiologic dataset: a comparison of random forests, boosted regression trees, conventional and penalized logistic regression for identifying independent factors associated with H1N1pdm influenza infections. BMC medical research methodology, 14(1), 99.

Boyacioglu, M. A., Kara, Y., & Baykan, Ö. K. (2009)." Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey". Expert Systems with Applications, 36(2), 3355-3366.

Farquad, M. A. H., Ravi, V., & Raju, S. B. (2014). "Churn prediction using comprehensible support vector machine: An analytical CRM application". Applied Soft Computing, vol. 19, pp. 31-40.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: the paradox of statistical learning. Trends in cognitive sciences, 19(3), 117-125.

Gorunescu, F. (2011, January). Introduction to Data Mining. In Data Mining (pp. 1-43). Springer Berlin Heidelberg.

He, B., Shi, Y., Wan, Q., & Zhao, X. (2014). "Prediction of Customer Attrition of Commercial Banks based on SVM Model". Procedia Computer Science, vol. 31, pp. 423-430.

Islam, T., Srivastava, P. K., Dai, Q., Gupta, M., & Zhuo, L. (2015). Rain Rate Retrieval Algorithm for Conical-Scanning Microwave Imagers Aided by Random Forest, RReliefF, and Multivariate Adaptive Regression Splines (RAMARS).Sensors Journal, IEEE, 15(4), 2186-2193.

J. Elith, J. R. Leathwick and T. Hastie, (2008),." A working guide to boosted regression trees", Journal of Animal Ecology. Aertsen, W., Kint, V., De Vos, B., Deckers, J., Van Orshoven, J., & Muys, B. (2012). Predicting forest site productivity in temperate lowland from forest floor, soil and litterfall characteristics using boosted regression trees. Plant and soil,354(1-2), 157-172.

Jiawei Han and Micheline Kamber, (2013), "Data Mining: Concepts and Techniques", 3th Edition, ISBN 978-0-12-381479-1, Elsevier..

Jin, C., & Jin, S. W. (2014). Software reliability prediction model based on support vector regression with improved estimation of distribution algorithms.Applied Soft Computing, 15, 113-120.

John W., (2014), "Data Smart: Using Data Science to transform information into Insight", ISBN: 978-1-118-66146-8, Wiley Publishing.

Kass, G. V. (1980),. An Exploratory Technique for Investigating Large Quantities of Categorical Data, Applied Statistics, Vol. 29, No. 2.

Kursa, M. B. (2014). Robustness of Random Forest-based gene selection methods. BMC bioinformatics, 15(1), 8.

Lin, F., Yeh, C. C., & Lee, M. Y. (2011). "The use of hybrid manifold learning and support vector machines in the prediction of business failure". Knowledge-Based Systems, 24(1), 95-101.

Lu, C. J., Lee, T. S., & Lian, C. M. (2012). "Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks". Decision Support Systems, 54(1), 584-596.

Max Kuhn and Kjell Johnson (2013).Applied Predictive Modeling. ISBN 978-1-4614-6849-3, Springer.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, A tutorial. Frontiers in neuro robotics, 7.

Roman Timofeev, (2004), "Classification and Regression Trees (CART) Theory and Applications", Master thesis, Humboldt University, Berlin.

Shehory, O., & Sturm, A. (2014). A Brief Introduction to Agents. InAgent-Oriented Software Engineering (pp. 3-11). Springer Berlin Heidelberg.

StatSoft electronic statistics textbook, 2010, http://www.statsoft.com/textbook .

Sut, N., & Simsek, O. (2011). Comparison of regression tree data mining methods for prediction of mortality in head injury. Expert systems with applications, 38(12), 15534-15539.

Ticknor, J. L. (2013). "A Bayesian regularized artificial neural network for stock market forecasting". Expert Systems with Applications, 40(14), 5501-5506.