Improved Naive Bayes Classification for Joint Investment Plan

MUFDA JAMEEL ALRAWASHDEH
Department of Mathematics
Qassim University, Buraidah,
SAUDI ARABIA

Abstract: - Large scale investments are mostly done by joint investors in different countries. Most of these investments involve collaboration with financial institutes of different countries. As the aspiration of governments to development their countries, they encourage investments. Financial institutes, at the same time, will set a guideline to decide with whom they will share the investment and collaborate based on profit maximization target. In this paper we are considering individual investors to collaborate with the financial institutes. Naïve Bayes is an ideal approach to aid the approval or rejection of this collaboration by the decision maker. The approach assumes independencies among the variables. However, this assumption may not always be realistic. Hence, this paper uses a method to improve the accuracy of Naïve Bayes approach by using a learning structure of feature variables in the model and apply it to joint investment plan applications. The introduction and use of new applied problem is not only helpful to show the application of the field but also attract researchers from social science to apply and use Bayes based methods which in turn contribute the development of the field with new insights.

Key-Words: - Joint Investment; Classification; Naïve Bayes Classifier; Structure Learning.

Received: March 27, 2021. Revised: December 19, 2021. Accepted: January 14, 2021. Published: February 8, 2022.

1 Introduction

With globalization in trades and investment, investors become interested in investing in different countries based on the willingness of governments and the collaboration of financial institutes to support the investment. For a country, the investment from abroad or within the country will bring huge advantages including job opportunities and development of the country. Investment destination countries, which are favorite for investors due to their investment rules and resources, may attract many joint investment proposals and hence decision making on approving or rejecting the proposal falls as a decision making problem by the financial institutes. Due to the uncertainty involved statistical approaches coined with artificial intelligence methods become useful in this regard. Data mining is the application of machine learning method to large databases [3]. A large volume of data is extracted to construct a simple model with a meaningful use, such as to provide high prediction accuracy by identifying patterns, anomalies and correlations. This application had been used in many fields such as in finance [13], manufacturing [10], medicine [15] and telecommunication [19]. Data mining can basically be classified into two types, which are predictive and descriptive models. Data mining builds models from data, using tools that vary both by the type of model built and by the type of algorithm used. In this study, predictive model is used to analyze the data. A model that predicts the value of a particular attribute and relates to class membership is called classification model or simply a classifier [9]. Classification is a data mining technique that assigns items in a collection to target categories or classes. The use of classification allows a good prediction of the target class for each given case in the data. It can, for example, be used to identify different level of classes of list of investment proposals as high profit, medium and low profit in our context. Classification is a task that can be used to identify the class labels for instances based on a set of features or attributes. A classification task usually begins with a data set in which the class assignments are known. For instance, a classification model that is used to predict investment risk can be developed based on observed data for many investment applicants over a time period. In addition to the past investment rating, the data might pursue company history, property ownership, years of existence, number and type of existing and past investments, and so on. The rating would be the target, the other attributes would be the predictors, and the data for each company would constitute a case. The properties of classification are discrete and do not imply order. A numerical target would be indicated from the continuous, floating-

point values. The simplest type of classification problem is binary classification. In a binary classification, the target attribute has only two possible outcomes, that is, approved or not approved. Multiclass targets have more than two outcomes: for example, low, medium, high, or unknown rating [9]. The Bayesian Network is a representation of probabilistic knowledge. It is also known as Bayes Network or directed acyclic graphical model. Bayesian network is a probabilistic graphical model that is used to represent a set of random variables and their conditional dependencies via a directed acyclic graph. A Bayesian network can be denoted by N(G, P). G is a directed acyclic graph, whose nodes are random variables X = {X1, X2, ..., Xn} and P is the set of conditional probability tables associated with each Xi [7, 12]. A special case of the Bayesian Network is denoted Naïve Bayes. Due to the simplicity of Naïve Bayes structure, it is easy to implement, and is thus appealing, provided that it gives good results.
Simplicity and speed make Naïve Bayes an ideal exploratory tool. It provides a flexible way for dealing with quantitative and discrete data, besides allowing any number of attributes or classes. Compare with other predictive techniques such as decision tree and neural network, Naïve Bayes is an asymptotically fast and space efficient tool that examines all its input [26]. Furthermore, small amount of bad data do not perturb the result significantly. It is not sensitive to irrelevant features. This could be good when handling many dumb variables [22]. Naïve Bayes has become a popular tool in data mining and other applications due to its simple, fast and easy nature for implementation [27].

Due to frequent uncertainty in the real world data set, misclassification patterns from the input samples generally restricted the utility of the classifier built [11]. The main problem confronting naive Bayes is its attribute conditional independence assumption. Therefore, learning the structural of the attributes in Naïve Bayes relationship is unavoidable. In this paper, study the interrelationship of the variables and construct a model based on the relationship that we observed from a hill-climbing algorithm. Hill-climbing (HC) is one of the available score-based learning algorithms by using directed arcs to explicitly represent variable dependencies [8]. The discussed method then will be used in the joint investment approval application. Investment is an important part of trade in a country's economy.
Using mathematical, statistical and computational methods produce a viable result for decision makers in the sector. The introduction of these approaches to solve problems in economics, trade and negotiation produce a new avenue of applications. This paper focuses on using statistical tools in proposing to solve an investment problem which can be further studied and used in real problems.

2 Background

International Investment is one of the investment strategies in which an investor diversifies his portfolio by purchasing various financial Instruments like shares, mutual funds, etc. or investing to acquire ownership or collaboration in different companies across the globe in order to maximize the return. In some countries, state owned industries may open up for international investors for collaborative investment with the government. In other cases financial institutes will try to attract investors local as well as international for joint investment with the aim of maximizing their profit.
With many proposals for this joint investment options, they need to select the promising collaboration based on the specification and experience of the investor partner with the aim of maximizing the expected return. Upon the call from the financial institute or under a normal application process, a bank or a financial institute will list and classify application based on some classification of each application by their details as investment capacity, years of existence in the business average education level of the management and so on. Conditional probabilities of scenarios will then be studied to approve or select best investment options with promising return. Bayesian network is an ideal approach for this kind of applications based on conditional probabilities. It has been used in different applications including in finance [20], medical applications [18], transportation industry [21], energy sector [5], water studies [12] and agriculture [6]. It is found to be a competitive option when compared to neural network [24], credit scoring model [14] and principal component analysis [25]. With uncertainty involved in financial institutes, the use of Bayesian networks becomes vital in financial applications. Naïve Bayes classifier is widely used in the finance market, such as bankruptcy prediction [1], credit risk assessment [4], business failure prediction [16] and so on.
This research, on the other hand, studies the application in joint investment approval. It further attempts and show that a learning in between independent nodes as assumed by Naïve Bayes classifier approach produce a better result.

3 The Proposed Approach

3.1 Data specification

This study use a score based structural learning algorithm in Bayesian Network to explore the dependencies among the variables. Based on this, it is aimed to show that with the structure learned among the attributes or variables in classification data set we can have higher classification accuracy than the Naïve Bayes classifier. We test the improvement in the classification of the data that we calculate the 10-fold cross validation. A data for this study was synthesized based on an initial set of data obtained from a financial firm (as given in Table (1) and (2)). It is highlighted that a more complete data can be used to test the proposed approach further as a future work.

3.2 Design of Naïve Bayes Model

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

where A and B are events and P(B)≠ 0. Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as evidence. P(A) is the priori of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance (here, it is event B). P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen. In simple terms, a Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. In classification learning, each instance in described by a vector of attributes values. Training data is a set of instances providing the known class.

Then, we can predict the class of the test instance based on evidence provided from the training data. From Fig (1), note that X1, X2, ..,Xn are independent of each other. For Naïve Bayes, we assume that n = 13 and all the 13 variables of our data are independent of each other. Since Naïve Bayes ignores the dependency among the variables, this affects the accuracy of Naïve Bayesian when the dependency among the variables is strong. Hence, we use the structural learning for the data to study the dependency among the feature variables.

3.2.1 Training

Let X be a training set of n records, C be the corresponding class and C = (c1, c2, . . . , cm) where m is the number of possibilities in class C be represented as (X,C) where X =(x1, x2, . . . , xn). Bayes classifier is a hybrid parameter probability model in essence:

$$P(C|X) = \frac{p(C)p(X|C)}{p(X)} \quad (2)$$

Table 1. Information and Variables of the Data

Data	Investment applications
Type of data	Discrete
Sample size	111 sets
Missing value	No
Class	Yes/No
Attributes	1. Age of the applicant company 2. Average education background of the management 3. Type of business of the company 4. Year in the proposed business of the company 5. Type of investment 6. Income of the company 7. Debt of the company 8. Financial institute contribution amount 9. Amount of income tax 10. Revenue of the company 11. Relationship between the company and the financial institute 12. Capital of the company 13. Size of the company

Table 2. Description of the data set (based on Table (1))

	Variables	Possible values	Description
1	Age	4	grouped by age: 1 representing new and 4 old
2	Education level	3	high school, undergraduate and graduate
3	Business type	4	partnership, proprietor, private, other
4	Year in business	6	grouped by year in the proposed business
5	Investment type	2	long ad short
6	Income	7	very low, low, moderate low, medium, high, moderate high, very high
7	Debt	7	none, low, moderate low, medium, moderate high, high, very high
8	Contribution	6	low, moderate low, medium, moderate high, high
9	Income tax	7	very low, low, moderate low, medium, moderate high, high , very high
10	Revenue	7	very low, low, moderate low, medium, moderate high, high , very high
11	Relationship	3	no relationship, exclusive, and non-exclusive
12	Capital	3	low, medium, high
13	Size	4	no relationship, exclusive, and non-exclusive

where P(C) is prior probability of the appearing probability of class C, P(X) is the probability from the observations, P(X|C) is the distribution probability of X in classes space. Assume the components xi of X are independent of each other. Since P(X|C) conditional probability cannot be computed directly in practice. Thus,

$$P(X|C) = \prod_i P(x_i|C) \quad (3)$$

3.2.2 Prediction

We make the prediction based on the evidence provided from the training data, calculating the posterior probability which is the final computed beliefs after the evidences have been propagated through the Bayes network. The posterior

probability, P(C|X) is calculated as given in Eq. (4) ([28]):

$$P(C|X) = \frac{P(C)\prod_i P(x_i|C)}{P(X)} \quad (4)$$

3.3 Design of Proposed Method

The proposed method has added relationship by using Bayesian network structural learning. The task of structure learning for Bayesian networks refers to learning the structure of the directed acyclic graph (DAG) from data. In other words, it determines which variables influence each other, establishing the dependency of variables (through arcs). Learning Bayesian network is an NP-hard problem and hence heuristic search is taken as an advantageous approach. Any heuristic based approach can be used for this purpose, including genetic algorithm, particle swarm optimization and so on. However, the hill climbing method is also found to be computationally effective and easier in terms of implementation.

It is also one the popular approach in this regard [8]. Hence, in this study, we use the Hill-Climbing (HC) algorithm to induce search on the space of the possible network. By using *'bnlearn'* package in R language to help implement HC algorithm [23], it can be determined whether there are relationships among the variables. If there are some relationships, then arcs will be added to show the dependencies as shown in Fig. (2). From the Fig. (2), observed that there are few additional arcs between nodes X_1&X_3; nodes X_2&X_n and nodes X_3&X_n. This shows that the presence of dependencies among these variables.

3.4 Cross-validation

A popular way to evaluate and compare models is on their ability to make predictions for "out-of-sample data", that is, future or unseen observations, using what we learned from the observed data. We can use cross-validation to test which of the models under consideration is able to learn the most from our data in order to make the better predictions.

Hence, the last step of our analysis is the cross-validation. Cross-validation involves using a subset of a partition of the data as a test data set. The remaining data are used to learn or train a model and the test data set is used to validate the model [17]. In order to get a stable estimate, we use 10-folds cross-validation. 10-fold cross validation performs the fitting procedure a total of ten times, with each fit being performed on a training set consisting of 90% of the total training set selected at random, with the remaining 10% used as a hold out set for validation. To do so, the data is first segmented into 10 equally (or almost equally) sized segments or folds. Here, the data set is repeatedly partitioned into two non-overlapping parts; a training set and a hold-out set. For each partitioning, the one fold of hold-out set is used for testing, while the remainder 9 folds are used for training as shown in the Fig. (3). Each subset was used once in the test set and nine times in the training set [2]. In the process of cross-validation, one fold of data is trained and tested for each partition and the accuracies averaged. Errors obtained from all the partitions are averaged. So, we can compare these two methods in term of the error estimates.

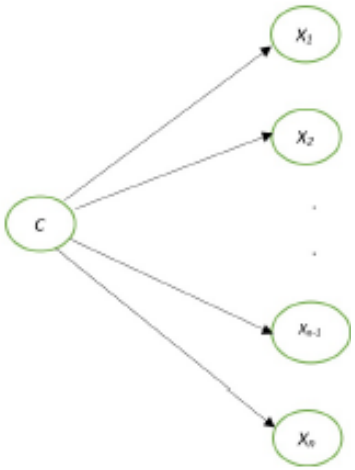

Fig. 1: An Illustration of the Naïve Bayes network structure

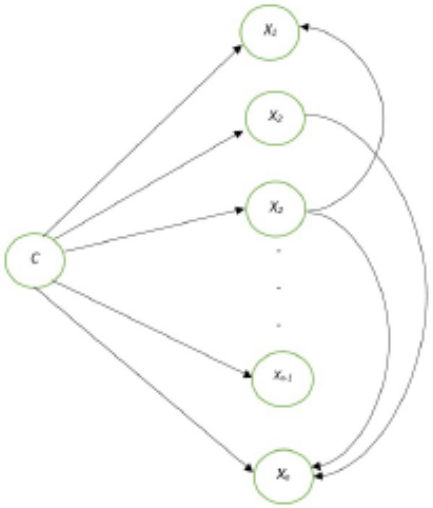

Fig. 2: A Naïve Bayes network structure with added relationships after structural learning

4 Results and Discussion

4.1 Naïve Bayes model

The structure of the Naïve Bayes network is shown in Fig. (4). From the graph, we can clearly see that

class node directly points to the 13 attributes. These 13 variables are conditionally independent. There is no relationship among the 13 variables. Naïve Bayes network has the strong assumption that each of the variables is not affected by another given the approval of investment is known. Any changes in each variable do not affect other parameters.

4.1.1 Performance Metric of the Naïve Bayes

Considering a binary classification problem with the 2 outcomes that is 'Approved' represented by 'Yes' and 'Reject' represented by 'No', a 2x2 confusion or error matrix is obtained from the framework, as given in Table (3). There are total of 111 instances in the matrix. Proportions that are properly classified is $\frac{(29+61)}{111}$ = 0.8108. This indicates that the accuracy for the performance of the Naïve Bayes is $\frac{(29+61)}{111}$ × 100% = 81.08%.

4.2 Structural Learning Model (The Proposed Method)

From Fig. (5), we can observe that there is no undirected arc shown in this graph. There are more arcs among variables compared with Fig. (4) after relationship is added into the model. This implies that there exists more than one relationship among variables.

From the graph, we know that capital of the financial institute is influenced by the type of investment and the investor's debt. Besides, relationship between investors and financial institutes is easily influenced by the investment amount they have proposed. Size of the company is dependent on 3 reasons that are capital provided, income of the applicant and income tax.

4.2.1 Performance Metric of the Proposed Method

From Table (4), the performance of the proposed method can be seen for the given 111 instances. For the Naïve Bayes with added relationship, the proportion that is properly classified is $\frac{(29+70)}{111}$ = 0.8919.

This indicates that the accuracy for the performance of the Naïve Bayes is $\frac{(29+70)}{111}$ 100% = 89.19%. Compared with the accuracy percentage in Table (3), there is an improvement of 8.11% for the performance of Naïve Bayes method.

4.3 Cross-Validation for Naïve Bayes and Proposed Method

Cross-validation used here is one the validation techniques, where it partitions the data into 10 stratified folds, performing the testing on one fold, while training on the other nine. This process is repeated 10 times and finally the average accuracy across the ten runs will be obtained. The proportions and percentage of both methods are compared in Table (5).

4.4 Discussion

In the study, a comparison between the Naïve Bayes and the proposed method-an improved Naïve Bayes was done. We analyze the data by using Naïve Bayes model and obtain the classification accuracy based on its assumption that the feature variables are independent of one another. However, in most real cases, this assumption is not realistic. Therefore, a hill-climbing algorithm is used to study the interrelationship among the variables, with that we establish an improved Naïve Bayes model. We improved the Naïve Bayes by learning the structure and the relationship among the feature variables. Using the bnlearn package

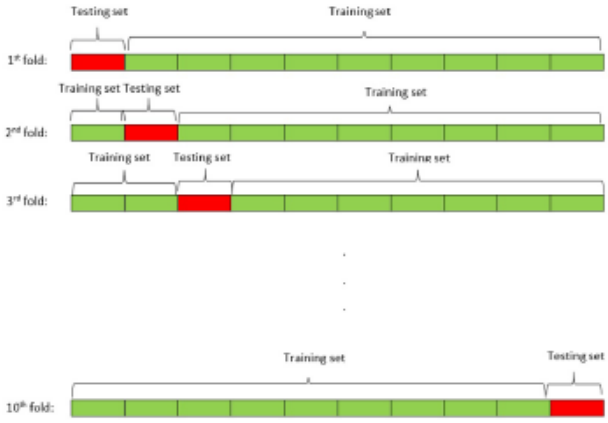

Fig. 3: An Illustration of the 10 fold cross-validation

Table 3. Confusion matrix table of the prediction by using Naïve Bayes model

	Yes	No
Yes	29	2
No	19	61

Table 4. The matrix table of the prediction by using Naïve Bayes with added relationship

	Yes	No
Yes	29	2
No	10	70

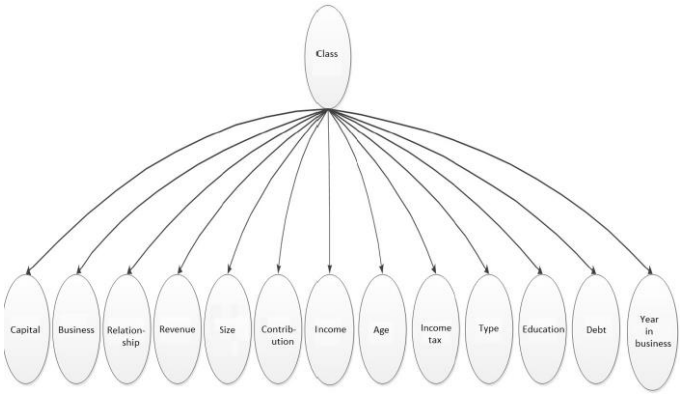

Fig. 4: Naïve Bayes Structure for the 13 variables

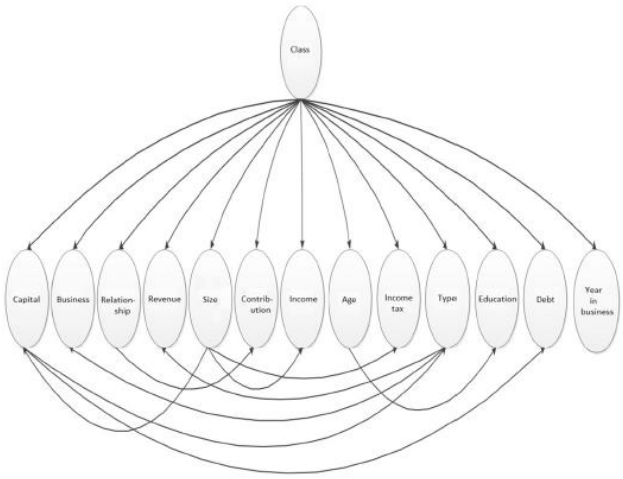

Fig. 5: The Naïve Bayes network structure with added relationship after structural learning using the hill-climbing algorithm on the features from the data set in R-language, we run the analysis for both Naïve Bayes and the improved model.

Table 5. Cross-validation tables for Naïve Bayes and Naïve Bayes with added relationship

Data segment	Naïve Bayes	Proposed method
1 - 11	9/11	9/11
12 - 22	11/11	11/11
23 - 33	6/11	7/11
34 - 44	6/11	7/11
45 - 55	10/11	11/11
56 - 66	10/11	11/11
67 - 77	6/11	10/11
78 - 88	10/11	11/11
89 - 99	11/11	11/11
100 - 111	9/11	8/11
Average	0.7932	0.8667
Percentage (%)	79.32	86.67

Table 6. Comparison of the accuracy percentage between Naïve Bayes and the proposed method

Method	Accuracy	10-fold Cross-validation
Naïve Bayes	81.08%	79.32%
The proposed method	89.19%	86.67%

A set of data consisting of 111 samples with 13 variables is used for demonstrating the proposed method. Continuous data has been discretized. This process is usually carried out as a first step towards making them suitable for numerical evaluation and implementation on digital computers. The results show that, the improved Naïve Bayes is better than Naïve Bayes in terms of classification accuracy. For Naïve Bayes, the classification accuracy is 81.08%. However, the classification accuracy is higher for the improved Naïve Bayes which is 89.19%. Furthermore, we run a 10-fold cross-validation for Naïve Bayes model and improved Naïve Bayes model. Hence, the results how that the improved Naïve Bayes performs better when compared to the Naïve Bayes method.

5 Conclusion and Future works

Naïve Bayes is a very useful method applied in different applications including investment approval by financial institutes. Since it showed a good performance for bank prediction.

It can also be applied and used for similar problems. A better prediction system may help governments and financial institutes to make decision. Its application is not limited to financial related issues but can be expanded and used in different fields where uncertainty exists. In this study a partially synthesized data is used to demonstrate the proposed approach. The result shows that finding interdependency of variables which normally are assumed to be independent in the Naïve Bayes method, actually produce a better accuracy.

As a future work, it is aimed to use a complete and more complex data for testing the proposed approach. Competition within multiple financial institutes is another future direction to consider as investors may apply to multiple collaborative investments and decide based on the outcome of their applications. Hence, game theoretic model can also be coupled with the Bayesian prediction proposed in this research. Applications in other domain of problems are also another future direction to explore more.

References:

[1] Aghaie A., Saeedi A., Using Bayesian Networks for Bankruptcy Prediction: Empirical Evidence from Iranian Companies, Proceeding of the 2009 International Conference on Information Management and Engineering, Kuala Lumpur, Maaysia, 3 - 5 April 2009, 450–455.

[2] Airola A., Pahikkala, T., Waegeman W., Baets, B. D., Salakoski, T., An experimental comparison of cross validation techniques for estimating the area under the ROC curve, J. Comput. Stat. & Data Anal. 55, 4, 2011, 1828–1844.
[3] Alpaydin E., Introduction to machine learning, MIT Press, USA, 2004, 54–55.
[4] Antonakis, A. C., Sfakianakis M. E., A_ ssessing Na¨ıve Bayes as a Method for Screening Credit Applicants, J. App. Stat. 36, 5, 2009, 537–545.
[5] Cinar, D., Kayakutlu G., Scenario analysis using Bayesian networks: A case study in energy sector, Knowledge-Based Systems, 23, 3, 2010, 267–276.
[6] Drury B., Valverde-Rebaza J., Moura M. F., de Andrade Lopes, A. A survey of the applications of Bayesian networks in agriculture, Eng. App. Artif. Intell., 65, 2017, 29–42.
[7] Friedman N., Geiger D., Goldszmidt M.,Bayesian network classifiers, Mach. learn. 29, 2, 1997, 131–163.
[8] Gamez J. A., Mateo J. L., Puerta J. M., Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood, Data Min. Knowl. Discov., 22, 1, 2011, 106–148.
[9] Hand D. J., Principles of data mining, Drug safety, 30, 7, 2007, 621–622.
[10] Harding J. A., Shahbaz M., Kusiak A., Data mining in manufacturing: a review, J. Manuf. Sci. Eng., 128, 4, 2006, 969–976.
[11] Hsieh N.C., Hung L. P., A data driven ensemble classifier for credit scoring analysis, J. Exp. Sys. Appl.: An Int. J. 37, 1, 2010, 534–545.
[12] Henriksen H. J., Rasmussen, P., Brandt G., Von Buelow D., and Jensen F. V., Public participation modelling using Bayesian networks in management of groundwater contamination, Env. Model. and Soft., 22, 8, 2007, 1101-1113.
[13] Kadam S., Raval, M., Data mining in finance, Int. J. Eng. Trends Technol 16, 2014, 377–381.
[14] Koh H. C., Tan W. C., Goh C. P., A Two-step Method to Construct Credit Scoring Models with Data Mining Techniques, Int. J. Bus. Inf., 1, 1, 2006, 96–118.
[15] Lavrac N., Zupan B., Data mining in medicine. In Data Mining and Knowledge Discovery Handbook-Springer, Boston, USA, 2005, 1107–1137.
[16] Li H., Sun, J., Wu J., Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods, Exp. Sys. with Appl., 37, 8, 2010, 5895–5904.
[17] Maunder M. N., Harley S. J., Using cross validation model selection to determine the shape of nonparametric selectivity curves in fisheries stock assessment models, Fisheries Res., 110, 2, 2011, 283–288.
[18] McLachlan S., Dube K., Hitman G. A., Fenton N. E., Kyrimi E., Bayesian networks in healthcare: Distribution by medical condition, Artif. Intell. Med., 107, 2020, 1–7.
[19] Nadaf M., Kadam V., Data mining in telecommunication, Int. J. Adv. Comput. Theory Eng., 2, 2013, 92–6.
[20] Neil M., Fenton, N. Using Bayesian networks to model the operational risk to information technology infrastructure in financial institutions, J. of Financ. Transform., 22, 2008, 131–138.
[21] Ni D., Leonard J. D.,Markov chain Monte Carlo multiple imputation using Bayesian networks for incomplete intelligent transportation systems data, Transp. Res. Rec., 1935, 1, 2005, 57–67.
[22] Ratanamahatana C. A., Gunopulos D., Scaling up the Naive Bayesian Classifier: Using Decision Trees for Feature Selection, Proceedings of Workshop on Data Cleaning and Preprocessing: in ICDM'02, Maebashi, Japan, December 9 - 12, 2002.
[23] Scutari M. (2021). Bnlearn - An R Package for Bayesian Network Learning and Inference, Available online: (accessed on 9th July 2021).
[24] Shorouq F. E., Saad G. Y., Applying Neural Networks for Loan Decisions in the Jordanian Commercial Banking System, J. Comput. Sci. Net. Secur. 10, 1, 2010, 209–214.
[25] Sustersic M., Mramor D., Zupanm J., Consumer credit scoring models with limited data, J. Exp. Sys., with Appl., 36, 3,2009, 4736–4744.
[26] Xhemali D., Hinde C. J., Stone R. G., Na¨ıve bayes vs. decision trees vs. neural networks in the classification of training web page. Inter. J. Comput. Sci., 4, 1, 2009, 16–23.
[27] Zhang C. L., Gui R. X., Yu Y., Zh, H. Y., Web-Scale Classification with Naive Bayes, Proceedings of the 18th International Conference on World Wide Web, Madrid, Spain, 20-24 April 2009, 1083–1084.
[28] Zhang W., Gao F., An Improvement to Naive Bayes for Text Classification, Procedia Engineering, 15, 2001, 2160–2164.

Conflict of Interest:

The author states no conflict of interest.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

Authors state no funding involved.