Journal of Autonomous Intelligence (2019) Volume 2 Issue 2.
doi:10.32629/jai.v2i2.40

Original Article

Classification of the Priority of Auditing XBRL Instance Documents with Fuzzy Support Vector Machines Algorithm

Guang Yih Sheu*

Associated Professor, Department of Accounting and Information System, Chang-Jung Christian University, Tainan, Taiwan, R.O.C..xsheu@mail.cjcu.edu.tw

ABSTRACT

Concluding the conformity of XBRL (eXtensible Business Reporting Language) instance documents law to the Benford's law yields different results before and after a company's financial distress. A new idea of applying the machine learning technique to redefine the way conventional auditors work is therefore proposed since the unacceptable conformity implies a large likelihood of a fraudulent document. Fuzzy support vector machines models are developed to implement such an idea. The dependent variable is a fuzzy variable quantifying the conformity of an XBRL instance document to the Benford's law; whereas, independent variables are financial ratios. The interval factor method is introduced to express the fuzziness in input data. It is found the range of a fuzzy support vector machines model is controlled by maximum and minimum dependent and independent variables. Therefore, defining any member function to describe the fuzziness in input data is unnecessary. The results of this study indicate that the price-to-book ratio versus equity ratio is suitable to classify the priority of auditing XBRL instance documents with the less than 30 % misclassification rate. In conclusion, the machine learning technique may be used to redefine the way conventional auditors work. This study provides the main evidence of applying a future project of training smart auditors.

Keywords: *Fuzzy support vector machines algorithm; Interval factor method; Benford's law XBRL; Audit*

ARTICLE INFO
Received: Apr 30, 2019
Accepted: June 5, 2019
Available online: June 18, 2019

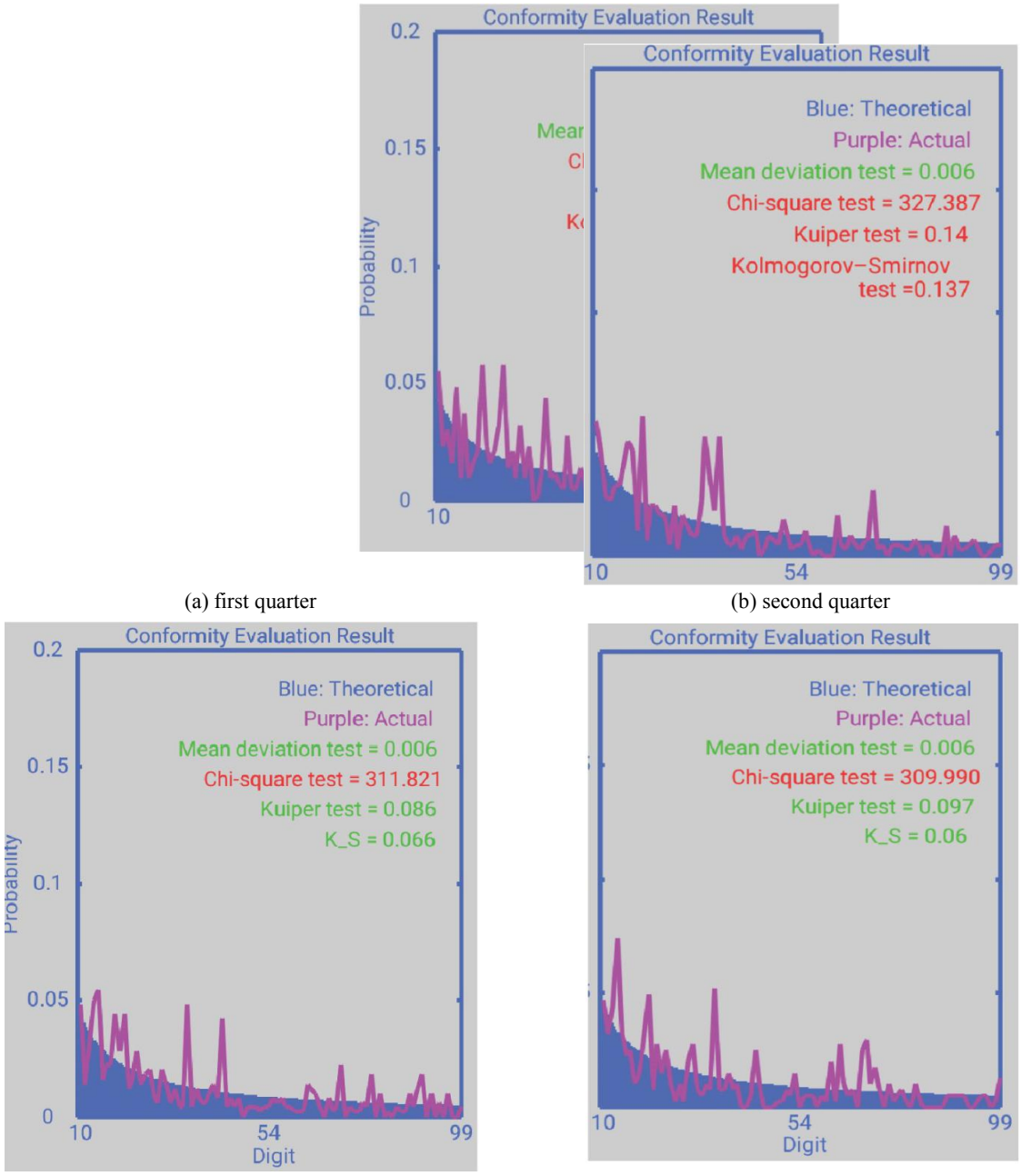
***CORRESPONDING AUTHOR**
Guang Yih Sheu, Associated Professor, Department of Accounting and Information System, Chang-Jung Christian University, Tainan, Taiwan, R.O.C; xsheu@mail.cjcu.edu.tw;

CITATION
Guang Yih Sheu. Classification of the Priority of Auditing XBRL Instance Documents with Fuzzy Support Vector Machines Algorithm. Journal of Autonomous Intelligence 2019; 2(2): 1-13. doi: 10.32629/jai.v2i2.40

1. Introduction

Detecting fraudulent documents is time-consuming if the total amount of documents to be detected is large. Any technique, which can help to quicken the detection of fraudulent documents, is welcomed. Benford's law[1] is one of such techniques. If digital probabilities calculated from a document doesn't obey the Benford's law[1], the unacceptable
conformity of this document to the Benford's law is concluded. Although this conclusion can be defeated by insufficient extracted digital data, the unacceptable conformity implies a larger likelihood of a fraudulent document.

In a previous project[2], an Android app named by aXBRL (audit of XBRL instance documents) was coded to evaluate the conformity of an XBRL instance document to the Benford's law[1]. An example of applying this aXBRL app is shown in **Figures 1(a)-1(d)**[3].



(a) first quarter (b) second quarter

(c) third quarter (d) fourth quarter

Figure 1. Audit of XBRL instance documents presented by the HOLUX company (http://www.holux.com/) (significance level = 0.01)

Figures 1(a)-1(d) show that apparently different results were obtained in checking XBRL instance documents presented by a HOLUX company (http://www.holux.com/) before and after the financial distress of this company. Attributing to this financial distress, the stock of HOLUX company was exchanged with full-cash delivery since 2017/11/20. In the meantime, evaluating the conformity of XBRL instance documents provided by this HOLUX company yields different results before and after this company's financial distress. In **Figure 1(b)**, only the mean deviation test statistic[4] indicates the acceptable conformity; whereas, only the Chi-squares test statistic[5] indicates the unacceptable conformity. The significance level is set to 0.01 in creating **Figures 1(a)-1(d)**. Accordingly, it seems that the HOLUX company rectified its financial condition for preventing its stock from being delisted from the stock market.

Figures 1(a)-1(d) result in an idea of applying the machine learning technique to help the detection of the fraudulent document since the unacceptable conformity concludes a large likelihood of detecting a fraudulent

document. The fuzzy support vector machines algorithm is chosen to implement such an idea. The dependent variable is a fuzzy variable quantifying the conformity of an XBRL instance document to Benford's law[1]. Creating this fuzzy variable instead of a deterministic one attributes to the experiences that inconsistent conformity of an XBRL instance document to the Benford's law is usually concluded.

Meanwhile, the independent variables are financial ratios including price-to-book and price-to-earnings ratios. Maximum and minimum values of these two ratios are available over a quarter of a year. Attributing to those price-to-book and price-to-earnings ratios and above-mentioned dependent variable, the classical support vector machines algorithm[6] is not used in this study.

The classical support vector machines algorithm was developed by[6]. It is now one of the popular classification methods. A support vector machine maps input data into a high-dimensional feature space. A hyperplane, which separates the input data into two classes, is next searched. This search continues until the optimal one, which maximizing the margins between two classes in the space, is found. A hyperplane is expressed in terms of linear or nonlinear kernels and few input data. These few input data are called support vectors. Maximizing the margins between two classes of input data is expressed as a quadratic programming problem. This quadratic problem is solved from its dual problem by introducing Lagrangian multipliers.

The classical support vector machine algorithm has been used to the stock market forecasting[7] and financial distress[8] or bankruptcy prediction[9]. As compared to the logistic and decision tree models, the most accurate classification is usually obtained using a classical support vector machines model. However, it has been concluded[10] that the classical support vector machine algorithm is sensitive to outliers or noises in training data. Therefore, the fuzzy support vector machine algorithm was developed to overcome this problem.

In existing fuzzy support vector machine models (e.g.[10]), separating fuzzy variables are created to simulate the fuzziness in dependent variable. Membership functions are next defined to describe the variation of these fuzzy variables. They are solved based on procedures of constructing classical support vector machines models.

Nevertheless, defining separating fuzzy variables to describe the fuzziness in dependent variables would not help matters in implementing the current study since the fuzziness in dependent variable of this study has not been clearly investigated. The reason of obtaining the inconsistent conformity of an XBRL instance document to Benford's law has not studied.

Fortunately, this study only intends to know the influence of fuzziness dependent variable on the range of hyperplane in a fuzzy support vector machines model. Therefore, the interval factor method[11] is applied to express the fuzziness in dependent variable. The fuzziness in independent variable is described in the same way. The independent and dependent variables are expressed in terms of interval variables times their mean (or midpoint) values. Ranges of hyperplanes in proposed fuzzy support vector machines models are expressed by maximum and minimum values of such interval variables and those mean values.

The interval factor method was developed to implement the interval finite element method. It was developed to solve the ranges of dependent variables. Available studies[11], which are related to this interval factor method, have no relations to the main interest of this study. The usual application of interval factor method is solving displacements caused by interval loads[11].

The remainder of this study is organized into five sections. In the next section, the aXBRL app is reviewed. In Section 3, the interval factor method is used to express the dependent and independent variables of this study. Section 4 presents the proposed fuzzy support vector machines model. In Section 5, the resulting fuzzy support vector machines model is applied to classify XBRL instance documents of listed companies at the Taiwan stock market. Based on this section, Section 6 presents the conclusion.

2. aXBRL APP

Benford's law is a theoretical relationship of digital probabilities in a document. The digital probability is equal to the total occurrence of a bin at such as leading

and first-two digits divided by the total occurrence of all bins at the same digits.

Suppose d_1, d_2....d_9 are the leading digital probabilities and d_{10}, d_{11}....d_{99} are the first-two digital probabilities. In addition, ξ_1, ξ_2.... ξ_9, ξ_{10}, ξ_{11}.... ξ_{99} are theoretical leading and first-two digital probabilities. They are equal to[1].

$$\xi_I=\log\left(1+\frac{1}{I}\right) \quad (1)$$

where I = 1, 2…99. The mean deviation (e.g.[4]), Chi-squares[5], Kuiper[13], and Kolmogorov-Smirnov test statistics[14,15] are used to quantify the difference between d_1, d_2....d_{99} and ξ_1, ξ_2.... ξ_{99}. This mean absolute deviation test statistic is equal to[4]

$$\text{mean deviation test statistic} = \frac{\sum_{i=n}^{N}|d_i-\xi_i|}{N-n+1} \quad (2)$$

where n = 1 and N = 9 for leading digital probabilities, n = 10 and N = 99 for first-two digital probabilities, | | is the absolute function. **Table 1**[12] lists the critical values for concluding mean absolute deviation test statistics.

Digits	**Range**	**Conclusion**
leading	0 to 0.006	close conformity
	0.006 to 0.012	acceptable conformity
	0.012 to 0.015	marginally acceptable conformity
	> 0.015	non-conformity
first-two	0 to 0.012	close conformity
	0.012 to 0.018	acceptable conformity
	0.018 to 0.022	marginally acceptable conformity
	> 0.022	non-conformity

Table 1. Conclusions of the mean absolute deviation test[12]

The Chi-square test statistic is[5]

$$\text{Chi-square test statistic} = \sum_{i=n}^{N}\frac{(Md_i-M\xi_i)^2}{M\xi_i} \quad (3)$$

where M is the total number of collected bins for computing digital probabilities. The higher calculated Chi-square test statistics, the more digital probabilities d_1, d_2....d_{99} deviate from ξ_1, ξ_2.... ξ_{99}.

Critical values for concluding Chi-square test statistics are computed from the inverse of right-tailed probability of the Chi-square distribution. These critical values can be calculated using the CHIINV function of the Excel. If p is the significance level and leading digital probabilities are used, it can be obtained: 13.362 (p = 0.1), 15.507 (p = 0.05), 20.09 (p = 0.01), and 26.124 (p = 0.001). If Chi-square test statistic above these outputs is calculated, the unacceptably conformity of d_1, d_2....d_{99} to the ξ_1, ξ_2.... ξ_{99} is concluded. This conclusion can be defeated with the probability of p. Similarly, if the first-two digital probabilities are calculated, it can be obtained: 106.649 (p = 0.1), 112.022 (p = 0.05), 122.942 (p = 0.01), and 135.978 (p = 0.001).

The Kuiper test statistic[13] is calculated based on cumulative probabilities; hence, digital probabilities d_1, d_2....d_{99} and ξ_1, ξ_2.... ξ_{99} are summed to obtain cumulative digital probabilities. It is derived:

$$\begin{aligned} D_1 &= d_1 & \Xi_1 &= \xi_1 \\ D_1 &= d_1 + d_2 & \Xi_1 &= \xi_1 + \xi_2 \\ &\vdots \\ D_9 &= \sum_{i=1}^{9} d_i & \Xi_9 &= \sum_{i=1}^{9} \xi_i \\ D_{10} &= d_{10} & \Xi_{10} &= \xi_{10} \\ D_{11} &= d_{10} + d_{11} & \Xi_{11} &= \xi_{10} + \xi_{11} \\ &\vdots \\ D_{99} &= \sum_{i=10}^{99} d_i & \Xi_{99} &= \sum_{i=10}^{99} \xi_i \end{aligned} \quad (4)$$

in which D and Ξ are the actual and theoretical cumulative digital probabilities; respectively. Based on Eq. (4), the Kuiper test statistic is equal to[13]

$$\text{Kuiper test statistic} = \max(D_i - \Xi_i) + \max(\Xi_i - D_i) \quad (5)$$

where i = 1, 2...9 (for d_1, d_2....d_9) or i = 10, 11…99 (for d_{10}, d_{11}....d_{99}) and max is the maximum function. The critical value for concluding Kuiper test statistics are estimated from the total number of bins extracted to compute digital probabilities. If over 100 bins have been employed to calculate digital probabilities, it was suggested[13]

$$\text{Critical values for concluding Kuiper test statistics} = \frac{K_U}{\sqrt{M}} \quad (6)$$

where K_U is equal to 1.62 (p = 0.1), 1.747 (p = 0.05), 2.001 (p = 0.01), and 2.303 (p = 0.001).

Similar to the Kuiper test statistic, the Kolmogorov-Smirnov test statistic is [14,15]

$$\text{Kolmogorov-Smirnov test statistic} = \max|D_i - \Xi_i| \quad (7)$$

where i = 1, 2...9 (for d_1, d_2....d_9) or i = 10, 11…99 (for d_{10}, d_{11}....d_{99}) and max is the maximum function. Suppose over 40 bins have been collected to compute the D_1, D_2..., D_{99}, the critical value for concluding Kolmogorov-Smirnov test statistics can be

estimated by[15]

$$\text{Critical values for concluding Kolmogorov} - \text{Smirnov test statistics} = \frac{K_S}{\sqrt{M}} \quad (8)$$

where K_S is equal to 1.22 (p = 0.1), 1.36 (p = 0.05), 1.63 (p = 0.01), and 1.95 (p = 0.001).

Previously, computing Eqs. (1)-(8) from XBRL instance documents requires the ACL[16], IDEA[17], and Excel. However, at Taiwan, owning the ACL and IDEA is too expensive. Parsing an XBRL instance document using the Excel requires the correct XBRL taxonomy. The XBRL taxonomy provides the structure of XBRL instance documents. But it is free to "extend" an XBRL taxonomy. Therefore, many XBRL taxonomies may be required in parsing numerous XBRL instance documents using Excel. Accordingly, an XBRL taxonomy impacts what can be completed with XBRL instance documents[18].

An alternative to the ACL, IDEA, and Excel may be the end-user programming. Hence, the aXBRL app[2] was coded to evaluate the conformity of an XBRL instance document to Benford's law. An XBRL taxonomy is not needed in the evaluation. The input data are the URL (uniform resource locator) of an XBRL instance document and significant level for concluding Chi-square, Kuiper, and Kolmogorov-Smirnov test statistics.

Figure 2 shows the GUI (graphical user interface) of aXBRL app. After inputting the required data, one of two buttons can be pressed to calculate leading or first-two actual digital probabilities. Titles of these two buttons are "use leading digits" and "use first-two digits"; respectively. For example, **Figure 2** was captured in calculating leading digital probabilities. The resulting digital probabilities are further used to calculate mean deviation, Chi-square, Kuiper, and Kolmogorov-Smirnov test statistics. The resulting test statistics are used to conclude the conformity of input XBRL instance document to the Benford's law.

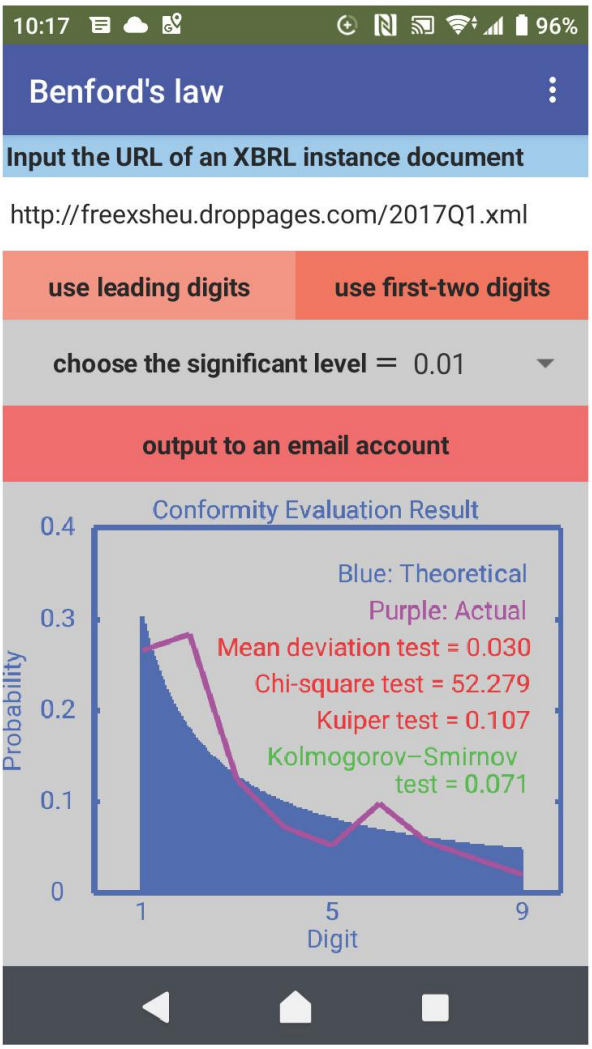


Figure 2. GUI of the aXBRL app

The resulting mean deviation, Chi-square, Kuiper, and Kolmogorov-Smirnov test statistics are presented in different colors; thus, the meanings of these test statistics can be visually understood. The meanings of these colors are explained below

1. Any test statistic in the red color denotes the unacceptable conformity of input XBRL instance document to the Benford's law.
2. Chi-square, Kuiper, and Kolmogorov-Smirnov test statistics in the green color denote the acceptable conformity of input XBRL instance document to the Benford's law.
3. The mean deviation test statistic in the green color represents the close conformity of input XBRL instance document to the Benford's law.
4. The mean deviation test statistic in the cyan color denotes the acceptable conformity of input XBRL

instance document to the Benford's law.

5. The mean deviation test statistic in the yellow color denotes the marginally acceptable conformity of input XBRL instance document to the Benford's law.

Figure 3 (in the next page) shows the flowchart of aXBRL app. Except for creating the GUI, the first step is sending a Get request to the input URL. The response contents are split into segments at chars "<" and ">". Digital data are found by testing whether each resulting segment can be converted into a number. Leading and first-two chars of each segment are extracted to calculate actual leading and first-two digital probabilities. The resulting leading and first-two digital probabilities are used to compute mean deviation, Chi-square, Kuiper, Kolmogorov-Smirnov test statistics. A separating subroutine is called to compute theoretical digital probabilities. The next step is plotting a graph to compare visually actual and theoretical digital probabilities and present the resulting mean deviation, Chi-square, Kuiper, Kolmogorov-Smirnov test statistics. The remaining step is sending the resulting graph to an e-mail account.

3. Interval Factor Method

Suppose y is a fuzzy variable quantifying the conformity of an XBRL instance document to the Benford's law and x is a financial ratio.

For simplicity, define the value of y within the range [-1,1]. Based on Section 2, assume that y = -1, $-\frac{1}{3}$-, $\frac{1}{3}$ -, and 1 to denote the unacceptable, marginally acceptable, acceptable, and close conformity of an XBRL instance document to the Benford's law; respectively. If the inconsistent conformity of an XBRL instance document to the Benford's law has been concluded, the resulting conformity is represented by an interval. For example, if the close to acceptable conformity is concluded using different test statistics, it is defined: $y = \left[\frac{1}{3},1\right]$.

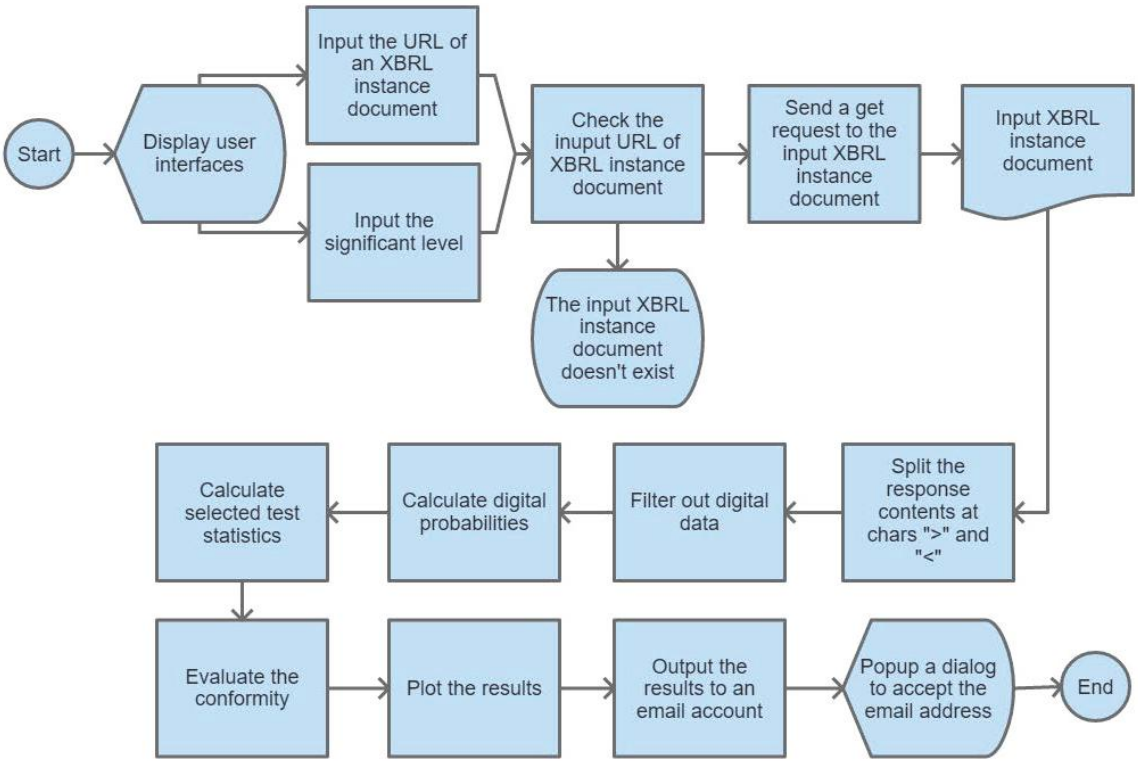


Figure 3. The flowchart of aXBRL app

Moreover, represent x and y by $x = \left[\underline{x},\bar{x}\right]$ and $y = \left[\underline{y},\bar{y}\right]$ in which the overline and underline denote minimum and maximum values; respectively. Further simplifying x and y yields

$$x = [x^C - \Delta x, x^C + \Delta x] \text{ and } y = [y^C - \Delta y, y^C + \Delta y] \quad (9)$$

where $x^c = \frac{\bar{x}+\underline{x}}{2}$, $\Delta x = \frac{\bar{x}-\underline{x}}{2}$, $y^c = \frac{\bar{y}+\underline{y}}{2}$, $\Delta y = \frac{\bar{y}-\underline{y}}{2}$, the superscript C and symbol Δ denote the mean (or

midpoint) value and maximum width of a variable; respectively.

Alternatively, Eq. (9) may be further modified to

$$x = x^C + [-\Delta x,\Delta x] \text{ and } y = y^C + [-\Delta y,\Delta y] \quad (10)$$

Extending Eqs. (9)-(10) to represent the range of a n × n matrix [A] results in

$$[A] = \left[\underline{[A]},\overline{[A]}\right] = [A]^C + (-\Delta[A],\Delta[A]) \quad (11)$$

where $[A]^c = \frac{\overline{[A]}+\underline{[A]}}{2}$ and $\Delta[A] = \frac{\overline{[A]}-\underline{[A]}}{2}$. Meanwhile, Eqs. (9) and (10) are further modified to [11]

$$x = \left[x^C\left(1-\frac{\Delta x}{x^C}\right),x^C\left(1+\frac{\Delta x}{x^C}\right)\right] = \left[1-\frac{\overline{x}-\underline{x}}{2x^C},1+\frac{\overline{x}-\underline{x}}{2x^C}\right]x^C = x^F x^C \quad (12)$$

$$y = \left[y^C\left(1-\frac{\Delta y}{y^C}\right),y^C\left(1+\frac{\Delta y}{y^C}\right)\right] = \left[1-\frac{\overline{y}-\underline{y}}{2y^C},1+\frac{\overline{y}-\underline{y}}{2y^C}\right]y^C = y^F y^C \quad (13)$$

where $x^F = \left[\underline{x}^F,\overline{x}^F\right]$, $y^F = \left[\underline{y}^F,\overline{y}^F\right]$, $\overline{x}^F = 1+\frac{\overline{x}-\underline{x}}{2x^C}$, $\overline{y}^F = 1+\frac{\overline{y}-\underline{y}}{2y^C}$, $\underline{x}^F = 1-\frac{\overline{x}-\underline{x}}{2x^C}$, $\underline{y}^F = 1-\frac{\overline{y}-\underline{y}}{2y^C}$, and the superscript F denotes the interval factor. In addition, it can be computed $(x^F)^C = 1$ and $\Delta x^F = \frac{\Delta x}{x^C}$; respectively. Similarly, it is derived: $(y^F)^C = 1$, $\Delta y^F = \frac{\Delta y}{y^C}$,

$$x^F = 1+\left[-\frac{\Delta x}{x^C},\frac{\Delta x}{x^C}\right] \text{ and } y^F = 1+\left[-\frac{\Delta y}{y^C},\frac{\Delta y}{y^C}\right] \quad (14)$$

In an attempt of operating Eqs. (9)-(14), the operation rules are additionally needed: Suppose $X = \left[\underline{X},\overline{X}\right]$, $Y = \left[\underline{Y},\overline{Y}\right]$, $X^F = \left[\underline{X}^F,\overline{X}^F\right]$, and $Y^F = \left[\underline{Y}^F,\overline{Y}^F\right]$; thus, X+Y, X-Y, XY, and $\frac{X}{Y}$ are calculated by[11]

$$X+Y = \left[\underline{X},\overline{X}\right]+\left[\underline{Y},\overline{Y}\right] = \left[\underline{X}+\underline{Y},\overline{X}+\overline{Y}\right] = \left[\underline{X}^F,\overline{X}^F\right]X^C+\left[\underline{Y}^F,\overline{Y}^F\right]Y^C = \left[\underline{X}^F X^C+\underline{Y}^F Y^C,\overline{X}^F X^C+\overline{Y}^F Y^C\right] \quad (15)$$

$$X-Y = \left[\underline{X},\overline{X}\right]-\left[\underline{Y},\overline{Y}\right] = \left[\underline{X}-\underline{Y},\overline{X}-\overline{Y}\right] = \left[\underline{X}^F,\overline{X}^F\right]X^C-\left[\underline{Y}^F,\overline{Y}^F\right]Y^C = \left[\underline{X}^F X^C-\underline{Y}^F Y^C,\overline{X}^F X^C-\overline{Y}^F Y^C\right] \quad (16)$$

$$XY = \left[\underline{X},\overline{X}\right]\left[\underline{Y},\overline{Y}\right] = \left[\min\left(\underline{XY},\underline{X}\overline{Y},\overline{X}\underline{Y},\overline{XY}\right),\max\left(\underline{XY},\underline{X}\overline{Y},\overline{X}\underline{Y},\overline{XY}\right)\right] = \left[\underline{X}^F,\overline{X}^F\right]X^C\left[\underline{Y}^F,\overline{Y}^F\right]Y^C = \left[\min\left(\underline{X}^F\underline{Y}^F,\underline{X}^F\overline{Y}^F,\overline{X}^F\underline{Y}^F,\overline{X}^F\underline{Y}^F\right),\max\left(\underline{X}^F\underline{Y}^F,\underline{X}^F\overline{Y}^F,\overline{X}^F\underline{Y}^F,\overline{X}^F\underline{Y}^F\right)\right]X^C Y^C \quad (17)$$

$$\frac{X}{Y} = \frac{\left[\underline{X},\overline{X}\right]}{\left[\underline{Y},\overline{Y}\right]} = \left[\underline{X},\overline{X}\right]\left[\frac{1}{\underline{Y}},\frac{1}{\overline{Y}}\right] = \frac{\left[\underline{X}^F,\overline{X}^F\right]X^C}{\left[\underline{Y}^F,\overline{Y}^F\right]Y^C} = \left[\underline{X}^F,\overline{X}^F\right]\left[\frac{1}{\overline{Y}^F},\frac{1}{\underline{Y}^F}\right]\frac{X^C}{Y^C} \quad (18)$$

in which max is the maximum function, min is the minimum function, $Y \neq 0$ should be satisfied in computing Eq. (18).

4. Fuzzy Support Vector Machines Model

Suppose N training data $(\mathbf{x_1},y_1),(\mathbf{x}_2,y_2),...,(\mathbf{x}_N,y_N)$ are collected in which $x = (x_1,x_2,...,x_n)$ is a vector of n financial ratios in the n dimensional space R^N. The y variable is our target variable.

Figure 4 illustrates the goal of constructing fuzzy support vector machines models[19] in which the linear classification is considered. Horizontal error bars in this figure indicates the fuzziness in x_i $(i = 1, 2,...,n)$; whereas, the fuzziness in the y variable is illustrated by vertical error bars. Therefore, the hyperplane may be adjusted to incorporate with the fuzziness in y and x_i. The hyperplane is defined by[6]

$$\mathbf{w}^T z + b = 0 \quad (19)$$

in which $\mathbf{w} = (w_1,w_2,...,w_n)$, $\mathbf{z} = [\phi(\mathbf{x_1}),\phi(\mathbf{x}_2),...,\phi(\mathbf{x}_n)]$, b is the offset, w_i $(i = 1, 2,...,n)$ is the weight, and ϕ is a mapping function from R^N to the feature space. Due to the fuzziness in x_i and y_i variables, the w_i and b vary within specific ranges.

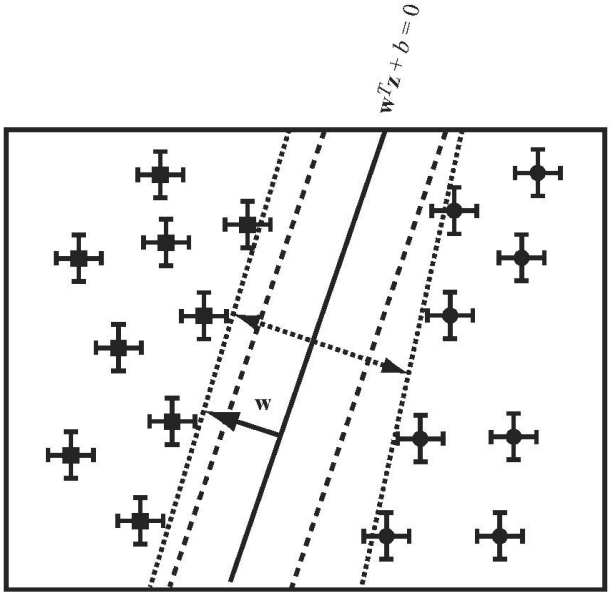


Figure 4. Development of a fuzzy support vector machines model

Eq. (19) indicates that $(\mathbf{x_1},y_1),(\mathbf{x}_2,y_2),...,(\mathbf{x}_N,y_N)$ are separated according to the function:

$$f(x) = \text{sign}(\mathbf{w}^T z + b) = \begin{cases} 1 & \text{if } y_i = 1 \\ -1 & \text{if } y_i = -1 \end{cases} \quad (20)$$

where sign is the sign function. If the

$(\mathbf{x_1},y_1),(\mathbf{x}_2,y_2),...,(\mathbf{x}_N,y_N)$ is linearly separable, each $(\mathbf{x_i},y_i)$ $(i = 1, 2,...,N)$ satisfies the inequalities:

$$\begin{aligned} \mathbf{w}^T z + b \geq 1 \qquad & \text{if } y_i = 1 \\ \mathbf{w}^T z + b \leq -1 \qquad & \text{if } y_i = -1 \end{aligned} \tag{21}$$

Eq. (21) is used to define a unique optimal hyperplane to classify linearly separable $(\mathbf{x_1},y_1),(\mathbf{x}_2,y_2),...,(\mathbf{x}_N,y_N)$ in which the margin between the projections of of two different classes is maximized. If $(\mathbf{x_1},y_1),(\mathbf{x}_2,y_2),...,(\mathbf{x}_N,y_N)$ are not linearly separable, new ξ_i $(i = 1, 2,...,N)$ variables are created to deal with classification violations. Eq. (21) is therefore modified to

$$y_i(\mathbf{w}^T z + b) \geq 1 - \xi_i \qquad i = 1, 2,...,N \tag{22}$$

Since these ξ_i $(i = 1, 2,...,N)$ variables are defined to deal with those x_i violating Eq. (19), $\sum_{i=1}^{N} \xi_i$ measures the misclassifications. Similar to the w_i and b variables, the ξ_i variables may be not deterministic.

The search of optimal hyperplane to classify $(\mathbf{x_1},y_1),(\mathbf{x}_2,y_2),...,(\mathbf{x}_N,y_N)$ can be expressed as the solution of problem[6,19]:

$$\text{minimize } \frac{1}{2}\mathbf{w}^T\mathbf{w} + \theta \sum_{i=1}^{N} \xi_i \text{ subjected to } y_i(\mathbf{w}^T z_i + b) \geq 1 - \xi_i \tag{23}$$

in which i = 1, 2...N and θ is a constant. This θ parameter may be considered as a regularization parameter. Turning this θ parameter can balance the margin maximization and classification violation in a support vector machines formulation. Detail discussions can be found in the reference.

Solving Eq. (23) is a quadratic programming problem. Such a quadratic programming problem is solved by minimizing a Lagrangian defined from this equation[6,19]:

$$L = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \theta \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \lambda_i [y_i(\mathbf{w}^T z_i + b) - 1 + \xi_i] - \sum_{i=1}^{N} \beta_i \xi_i \tag{24}$$

where $\lambda_i \geq 0$ and $\beta_i \geq 0$. In addition, Eq. (24) is not directly solved; whereas, it is transformed by the duality principle to solve the λ_i $(i = 1,2,...,N)$ variables[6,19]:

$$\text{minimize } W(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i\lambda_j y_i y_j \mathbf{z_i}^T\mathbf{z_I} \text{ subjected to } \sum_{j=1}^{N} \lambda_i y_i = 0 \tag{25}$$

where $\lambda = (\lambda_1,\lambda_2,...,\lambda_N)$. The next step is computing $\frac{\partial L}{\partial w}$, $\frac{\partial L}{\partial b}$, and $\frac{\partial L}{\partial \xi_i}$ $(i = 1, 2..., N)$ [6,19]. However, it is considered the fuzziness in y and x_j $(j = 1, 2..., n)$ variables; therefore, w_i, b, and ξ_i variables should range for incorporating with such fuzziness. Similarly manipulating Eqs. (12) and (13), it is defined:

$$w = (w^F)^T w^C = \left(w_1^F w_1^C, w_2^F w_2^C, ..., w_N^F w_N^C\right), \; b = b^F b^C, \text{ and } \xi_i = \xi_i^F \xi_i^C \tag{26}$$

where i = 1, 2..., N, $w^F = \left(w_1^F, w_2^F, ..., w_N^F\right)$, and $w^C = \left(w_1^C, w_2^C, ..., w_N^F w_N^C\right)$. Applying Eq. (26) and the chain rule to computing $\frac{\partial L}{\partial w}$, $\frac{\partial L}{\partial b}$, and $\frac{\partial L}{\partial \xi_i}$ (i = 1, 2…, N) derivatives yields

$$\begin{aligned} &\frac{\partial L}{\partial \mathbf{w}} = \frac{1}{2}\left[\frac{\partial\left(w_1^2\right)}{\partial\left(w_1^F w_1^C\right)}, \frac{\partial\left(w_2^2\right)}{\partial\left(w_2^F w_2^C\right)}, ..., \frac{\partial\left(w_N^2\right)}{\partial\left(w_N^F w_N^C\right)}\right] - \\ &\sum_{j=1}^{N} \lambda_i y_i \left[\frac{\partial\left(w_1^2\right)}{\partial\left(w_1^F w_1^C\right)}, \frac{\partial\left(w_2^2\right)}{\partial\left(w_2^F w_2^C\right)}, ..., \frac{\partial\left(w_N^2\right)}{\partial\left(w_N^F w_N^C\right)}\right]^T \mathbf{z_i} \\ &= (\mathbf{w}^F + \mathbf{w}^C)^T\mathbf{w} - (\mathbf{w}^F + \mathbf{w}^C)^T \sum_{i=1}^{N} \lambda_i y_i \mathbf{z_i} \\ &= 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{z_i} \end{aligned} \tag{27}$$

$$\frac{\partial L}{\partial \mathbf{b}} = \frac{\partial L}{\partial \mathbf{b}}\frac{\partial b}{\partial \mathbf{b^F}} + \frac{\partial L}{\partial \mathbf{b}}\frac{\partial b}{\partial \mathbf{b^C}} = -(\mathbf{b^F} + \mathbf{b^C}) \sum_{i=1}^{N} \lambda_i y_i = 0 \Rightarrow \sum_{i=1}^{N} \lambda_i y_i = 0 \tag{28}$$

$$\frac{\partial L}{\partial \xi_i} = \frac{\partial L}{\partial \xi_i}\frac{\partial \xi_i}{\partial \xi_i^F} + \frac{\partial L}{\partial \xi_i}\frac{\partial \xi_i}{\partial \xi_i^C} = -\theta\left(\xi_i^F + \xi_i^C\right) - \lambda_i\left(\xi_i^F + \xi_i^C\right) - \beta_i\left(\xi_i^F + \xi_i^C\right) = 0 \Rightarrow \theta = \lambda_i + \beta_i \tag{29}$$

Comparing Eqs. (27)-(29) to the derivation of classical support vector machines models[6,19] can find that they look similar except for the fuzziness in x_i and y_i (i = 1, 2..., n) variables. Further substituting Eq. (27) into Eq. (19) obtains

$$\mathbf{w}^T z + b = \sum_{i=1}^{N} \lambda_i y_i \mathbf{z}^T\mathbf{z_i} + b = \sum_{i=1}^{N} \lambda_i y_i K(\mathbf{z}.\mathbf{z_i}) + b = 0 \tag{30}$$

where K is the kernel function. In addition, Eq. (20) is modified to

$$f(x) = \text{sign}\left[\sum_{i=1}^{N} \lambda_i y_i K(\mathbf{z}.\mathbf{z_i}) + b\right] = \begin{cases} 1 & \text{if } y_i = 1 \\ -1 & \text{if } y_i = -1 \end{cases} \tag{31}$$

The linear kernel is chosen for showing that even a linear kernel can be used to provide useful help to the audit of XBRL instance documents. This linear kernel is defined by[6,19]

$$K(\mathbf{X},\mathbf{Y}) = \mathbf{X}^T\mathbf{Y} \tag{32}$$

where **X** and **Y** are two points. From Eq. (32), it is derived: $\mathbf{z} = \mathbf{x}$ and $\mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{z_i} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x_i}$.

Substituting these resulting expressions and Karush-Kuhn-Tucker condition to the constraints in Eq. (24) yields[6,19]

$$0 \le \lambda_i \le \theta,\ \beta_i \ge 0 \quad (33)$$

$$\lambda_i[y_i(\mathbf{w}^T\mathbf{z}_i + b) - 1 + \xi_i] = 0 \quad (34)$$

$$\beta_i\xi_i = 0 \quad (35)$$

in which i =1, 2..., N. Since Eq. (32) varies linearly with respect to different $\mathbf{x_i}$ or y_i (i = 1, 2…, N), Eqs. (27) and (33) indicate that $\overline{\mathbf{w}}$ and $\underline{\mathbf{w}}$ values are obtained from two of the following four combinations of maximum or minimum $\mathbf{x_i}$ and y_i (i = 1, 2..., N) values. The w, which are calculated from the other two combinations of maximum or minimum $\mathbf{x_i}$ and y_i (i = 1, 2..., N) values, are used to verify the resulting $\overline{\mathbf{w}}$ and $\underline{\mathbf{w}}$ values:

1. $\overline{\mathbf{x_i}}$ and $\overline{y_i}$ (or $\mathbf{x_i^C}\overline{\mathbf{x_i^F}}$ and $y_i^C\overline{y_i^F}$);
2. $\overline{\mathbf{x_i}}$ and $\underline{y_i}$ (or $\mathbf{x_i^C}\mathbf{x_i^F}$ and $y_i^C\underline{y_i^F}$);
3. $\underline{\mathbf{x_i}}$ and $\overline{y_i}$ (or $\mathbf{x_i^C}\mathbf{x_i^F}$ and $y_i^C\overline{y_i^F}$); and
4. $\underline{\mathbf{x_i}}$ and $\underline{y_i}$ (or $\mathbf{x_i^C}\overline{\mathbf{x_i^F}}$ and $y_i^C\underline{y_i^F}$);

Meanwhile, the sequential minimal optimization (SMO) algorithm[25] can be used to solve the λ_i (i = 1, 2,..., N) variables.

The point $\mathbf{x_i}$ (i = 1, 2..., N) with the condition $\lambda_i > 0$ is called a support vector. Two types of support vectors exist. The $\mathbf{x_i}$ with $\theta > \lambda_i > 0$ lies on the margin of the hyperplane. The other $\mathbf{x_i}$ with $\lambda_i = \theta$ is misclassified. The fuzziness in $\mathbf{x_i}$ and y variables affects the members of correctly classified and misclassified support vectors. This may be an important difference between the classical and fuzzy support vectors machines algorithms.

Conclusively, this study constructs a fuzzy support vector machines model with next five steps:

1. Suppose $(\mathbf{x_1},y_1),(\mathbf{x_2},y_2),...,(\mathbf{x_N},y_N)$ are the input data. Determine $(\overline{\mathbf{x_i}},\overline{y_i})$, and $(\underline{\mathbf{x_i}},\underline{y_i})$ values. This step can be completed using a spreadsheet software. If x_i is deterministic, $\overline{\mathbf{x_i}} = \underline{\mathbf{x_i}}$ is set. Similarly, $\overline{y_i} = \underline{y_i}$ is considered, if the y_i is deterministic.
2. Following the discussion below Eq. (35), construct four fuzzy support vector machines models.
3. Determine $\overline{\mathbf{w}}$, $\underline{\mathbf{w}}$, $\overline{b}$, $\underline{b}$, and ranges of hyperplanes according to the resulting fuzzy support vector machines models.
4. If verifying the resulting $\overline{\mathbf{w}}$, $\underline{\mathbf{w}}$, $\overline{b}$ and $\underline{b}$, values is desired, an interval Monte Carlo simulation is implemented. This interval Monte Carlo simulation is implemented by sampling the y and $\mathbf{x_i}$ (i = 1,2,..., N) variables between their maximum and minimum values. Use the resulting samples to construct classical support vector machines models and determine **w** and b values. Compare the resulting w and b values with the resulting $\overline{\mathbf{w}}$, $\underline{\mathbf{w}}$, $\overline{b}$ and $\underline{b}$ values in the third step.

The above four steps provide an alternative to existing methods (e.g. Lin and Wang, 2002) of constructing fuzzy support vector machines models. Different from those previous methods, sufficient data are unavailable in the succeeding study to define any membership function for representing the fuzziness in input data. However, implementing the current fuzzy support vector machines models requires only the existing software, which was used to construct classical support vector machines models. Although more computation is required to construct a fuzzy support vector machines model, this computation is acceptable. With such as R and Python programming languages, constructing a classical support vector machines model is easy.

5. Result and Discussion

Total of 132 XBRL instance documents is randomly chosen for the Taiwan stock market. These 132 XBRL instance documents are downloaded from the web site http://mops.twse.com.tw/mops/web/t203sb01. All the 132 XBRL instance documents were created over the third and four quarters of 2017. Among these 132 XBRL instance documents, 106 XBRL instance documents come from 52 randomly selected companies. These 52 companies didn't suffer from any financial distress. Other 26 (= 132-106) XBRL instance documents come from 13 randomly chosen companies with full-cash delivery stocks. They suffered from financial distresses. But the financial distress may not occur during third and fourth quarters of 2017.

Figure 5 (in the next page) illustrates the data processing procedures. The conformity of 132 XBRL instance documents to the Benford's law is first evaluated using the aXBRL app. First-two digital probabilities are used in this conformity evaluation. The required significance level is set to 0.01. The results are quantified by the y variable.

Meanwhile, data of financial ratios are obtained

from the TEJ database (http://www.finasia.biz/ensite/). Total 30 financial ratios are chosen. Table 2 lists these 30 financial ratios. They are classified into two types. Price-to-earnings and price-to-book ratios are classified into the uncertain type since maximum and minimum values of them are available over each quarter of 2017. Other financial ratios are classified as the deterministic type.

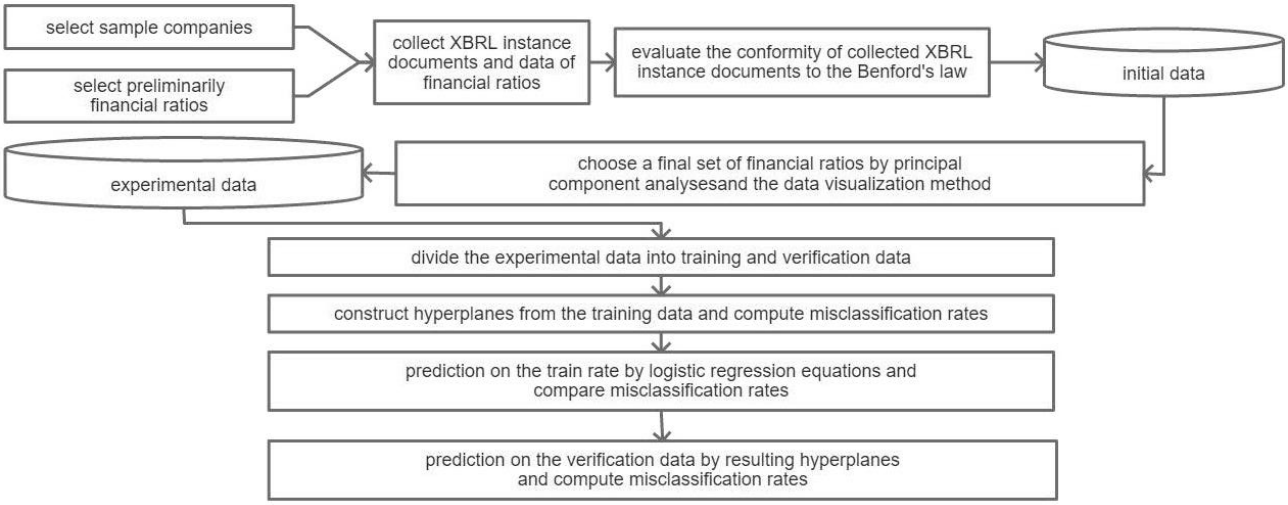


Figure 5. Data processing procedures

Type	Financial ratio
deterministic	price-to-book ratio, price-to-earning ratio
uncertain	return on assets, total assets turnover, quick ratio, gross profit margin, cash flow ratio, net profit margin, debt to equity ratio, ratio of shareholders' equity to fixed assets, earnings per share, equity ratio, total asset growth, operation income growth rate, gross margin growth rate, turnovers of account receivables, fixed assets turnover, operation income/capital, pretax income/capital, debt to total assets, debt to equity ratio, fixed assets/total assets, inventory turnover ratio, debt ratio, berry ratio, return on equity, current ratio, average collection days, growth rate of return on total assets, long term funds to fixed assets

Table 2. Selection of financial ratios (I)

Principal component analyses are next generated to provide clues to the selection of financial ratios for constructing fuzzy support vector machines models. **Figure 6** is the scree plot. Accordingly, it is a line plot of the eigenvalues of principal components in an analysis. A scree plot is used to determine the number of principal components to keep in principal component analysis. Since the price-to-earnings and price-to-book ratios are uncertain, maximum, mean, and minimum values of these financial ratios are separately used to implement three principal component analyses. However, the results are identical. Therefore, only one curve is plotted in **Figure 6**. From this figure, it is chosen five financial ratios most relevant to the first four principal components. **Table 3** lists the results.

Principal components	Financial ratio
first	equity, quick, and current ratios, return on assets, net profit margin
second	price-to-book ratio, pre-tax in-come/capital, earnings per share, operation income/capital, return on equity
third	net profit margin ratio of shareholders' equity to fixed assets, total asset growth, return on equity, return on assets
fourth	total asset turnover, turnovers of account receivables, inventory turnover ratio, cash flow ratio, fixed assets, turnover

Table 3. Selection of financial ratios (II)

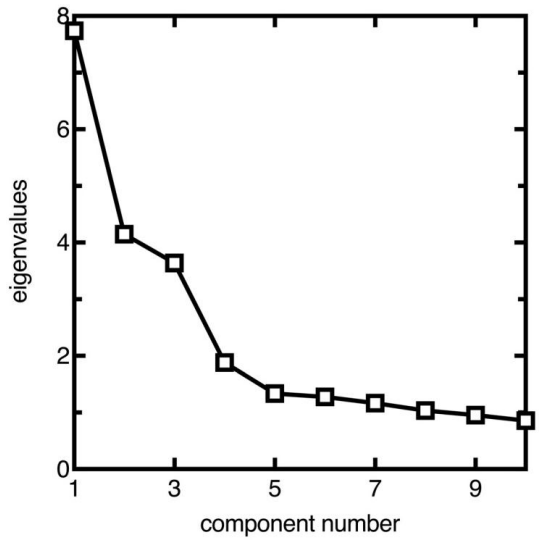


Figure 6. Scree plot

Scatter plot matrices are next created to confirm the selection of financial ratios in **Table 3**. **Figure 7** (in the next page) shows an example. This figure is plot based on XBRL instance documents, which was presented at the third quarter of 2017. Observing this figure can easily expect that adopting the earnings per share versus equity ratio is not suitable for building fuzzy support vector machines models. Similar inspections are implemented. The final result is the price-to-book ratio versus equity ratio to construct fuzzy support vector machines models since satisfactory misclassification rates are expected.

Since any XBRL instance document was created to present the latest financial information over a quarter of a year, the following discussion is separated according to the third and fourth quarters of 2017. In each quarter, the XBRL instance documents without any financial distress are used as training data to construct a fuzzy support vector machines model. Other XBRL instance documents are used as verification data.

Figures 8(a)-8(b) show the classification results for the training data. The θ parameter (in Eq. (23)) is set to 0.5. These classification results are represented by blue and red zones in these four figures. The color changes in these two figures represent the movement of hyperplanes due to the fuzziness in the price-to-book ratio and y variable. If a new data point locates at the blue zone in Figures 8(a)-8(b), it is considered the XBRL instance document represented by this data point is less possibly fraudulent. Whereas, if a new data point locates at the red zone in these four figures, a prior audit is suggested to the XBRL instance document.

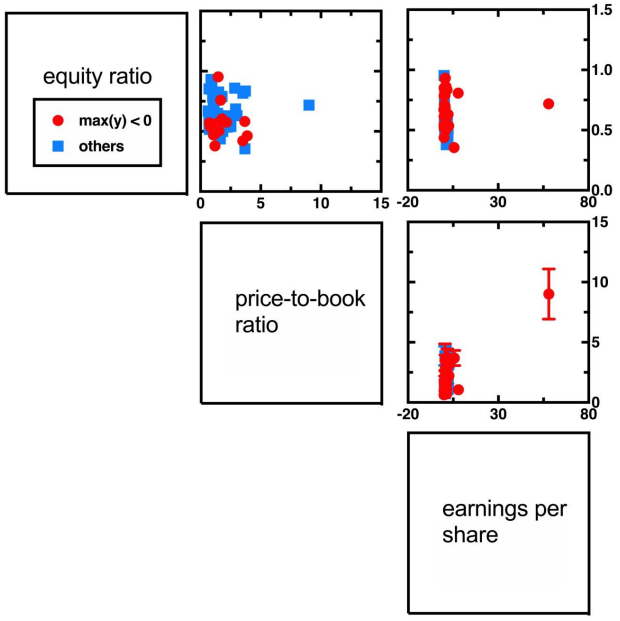


Figure 7. Selection of financial ratios (III)

The misclassification results obtained in creating Figures 8(a)-8(b) are listed in **Table 4**. It can be observed that the misclassification rates are less than 30 %. In addition, existing studies, which are devoted to the classical support vector machines algorithm, usually adopt different θ values to move the hyperplanes. This move of hyperplanes can improve the misclassification rates. However, observing Figures 8(a)-8(b) finds that it is difficult to change the θ value to improve the misclassification rate. Many blue or red points locate near the hyperplane. Moving the hyperplane is beneficial to the improvement of misclassification rates. Accordingly, this study doesn't move the hyperplane and inspect the corresponding misclassification rate.

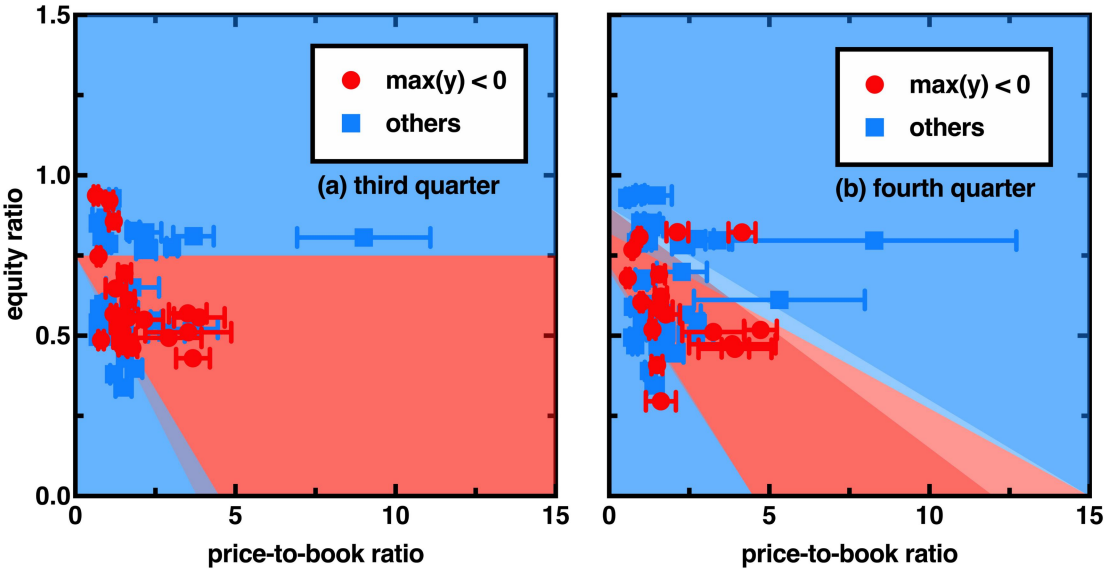


Figure 8. The search of more possibly fraudulent XBRL instance document

Quarter	max(y) < 0	others
third	20.6-27.5 %	20-23.3 %
fourth	13.5 %-16.2 %	18.9-21.6 %

Table 4. Misclassification rates obtained in creating Figures 8(a)-8(b)

Furthermore, the resulting fuzzy support vector machines models in Figures 8(a)-8(b) are applied to the verification data. **Table 5** lists the misclassification rate.

Quarter	max(y) < 0	others
third	25 %	25 %
fourth	0 %	9 %

Table 5. Misclassification rates obtained in verifying fuzzy support vector machines models shown in Figures 8(a)-8(b)

Observing **Table 5** finds that the classification results are still satisfactory, especially the misclassification rate in classifying the more possibly fraudulent XBRL instance documents (max(y) < 0). Some misclassification rates are even equal to 0 %. These results illustrate that the fuzzy support vector machines algorithm can be used to redefine the way conventional auditors work. Since the fuzzy support vector machines algorithm is one of the machine learning technique, the current section indicates a new application of the machine learning technique.

6. Conclusions

This study develops a new application of fuzzy support vector machines algorithm. Fuzzy support vector machines models are constructed to separate more possibly fraudulent XBRL instance documents from others. The dependent variable in these fuzzy support vector machines models is a fuzzy variable describing the inconsistent conformity of an XBRL instance document to the Benford's law (1938). The independent variables are the price-to-book and equity ratios. Auditors may use the resulting fuzzy support vector machines models to determine which XBRL instance document is audited first. The theoretical background of determining which XBRL instance document is more possibly fraudulent is Benford's law (1938).

It has been used the proposed fuzzy support vector machines models to classify XBRL instance documents, which were presented by companies with full-cash delivery stocks. The goal is finding more possibly fraudulent XBRL instance documents. The misclassification rate is less than 30 %.

This study demonstrates that the machine learning technique (e.g. the fuzzy support vector machines algorithm) can improve the way conventional auditors work. It will denote the main evidence of applying a future project of training smart auditors funded by the Taiwan's ministry of education.

Acknowledgements

This article is funded in part by the Ministry of Science and Technology of Taiwan, R.O.C., under Grant No. 106-2813-C-309-027-H5.

References

1. Benford, F. “The law of anomalous numbers”, Proceedings of the American Philosophical Society, Vol. 78, 1938.
2. Sheu Guang Yih. “aXBRL: Search of fraudulent XBRL instance documents with an Android app SoftwareX, Vol. 9, 2019.
3. Sheu Guang Yih, Chen, Y. X. “A Research of Integrating Fuzzy Support Vector Machines Model and an Android App to Assist the Audit of XBRL Instance Documents”, College Student Research Report, 107-2813-C-309-010-H, MOST, Taiwan, 2019 (in Chinese)
4. Hoaglin D. C., Mosteller F., Tukey J. W. “Understanding Robust and Exploratory Data Analysis”, John Wiley & Sons, 2008.
5. Greenwood. P. E., Nikulin, M. S. “A Guide to Chi-squared Testing”, John Wiley & Sons, 2008.
6. Vapnik, V. “Estimation of Dependences based on Empirical Data”, Springer-Verlag, 1982.
7. Park, M, Lee, M. L., Lee, J. “Predicting stock market indices using classification tools”, Asian Economic and Financial Review, 2019.
8. Tian, Y., Yang, W., Lia, G., Zhao, M. “Predicting non-life insurer's insolvency using non-kernel fuzzy quadratic surface support vector machines”, Journal of Industrial and Management Optimization, Vol. 15, 2019.
9. Cho, P., Chang, W., Song J.-W. “Application of instance-based entropy fuzzy support vector machine in peer-to-peer lending investment decision”, IEEE Access, 7, 2019.
10. Lin, C.-F. and Wang, S. D. “Fuzzy support vector machines”, IEEE Transaction on Neutral Networks, 2002.
11. Wei, G. “Interval finite element analysis using interval factor method”, Computational Mechanics, 2007.
12. Nigrini, M. J. and Wells, J. T. “Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection, New Jersey, USA: John Wiley & Sons, 2012.
13. Kuiper, N. H. “Tests concerning random points on a circle”, Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Series A., 1960.
14. Kolmogorov, A. “Sulla determination empirica di una legge di distribuzione”, Giornale dell’istituto italiano degli attuari, 1993
15. Smirnov, N. “Table for estimating the goodness of fit of empirical distributions”, Annals of Mathematical Statistics, 1948.
16. Hall, J. A. “Information Technology Auditing”, Cengage Learning, 2010.
17. Turnbull, C. S. “Fraud Investigation Using IDEA”, Ekaros Analytical Inc., 2003.
18. Hoffman, C. http:// http://xbrl.squarespace.com, 2010.
19. Kecman, V. “Learning and Soft Computing, Support Vector Machines, Neural Networks and Fuzzy Logic Models”, The MIT Press, 2001.
20. John, P. T. “Sequential Minimal Optimization: a Fast Algorithm for Training Support Vector Machines”, CiteSeerX 10.1.1.43.4376, 1998.