

A COMPUTER-IMPLEMENTED SEMI-SUPERVISED MACHINE LEARNING SYSTEM, METHOD AND COMPUTER READABLE MEDIUM FOR DETERMINING CORPORATE CREDIT RATING BASED ON A MULTI-FACTOR DISTANCE ALGORITHM

TECHNICAL FIELD

[0001] The present disclosure relates generally to a system, method and computer readable medium for determining a corporate credit rating, and more particularly to a computer-implemented semi-supervised machine learning system, method, and computer readable medium for determining a credit rating based on a multi-factor distance algorithm.

BACKGROUND

[0002] A corporate credit rating relates to the likelihood that a corporation will fully meet its financial obligations as they come due. A company's corporate credit rating indicates its relative ability to pay its creditors. From a borrower's perspective, a corporate credit rating is a formal, and independent opinion of a borrower's ability to service its debt obligations.

[0003] Expert subjective rating and statistical quantitative model are the two major aspects of international mainstream methodology in the field of corporate credit rating research. However, there are several problems in the existing solutions for determining corporate credit rating.

[0004] The expert subjective rating method selects certain factors for rating a company and decides the weight of the factors by anthropogenic method which are highly driven by manual influence. Such method requires years of subject matter experience from a specific industry and is used extensively by leading international independent rating agencies, for example, S&P, Moody's and Fitch. Statistical quantitative models such as Altman Z-Score, on the other hand, require a comprehensive financial standard and rich financial data. Most of the domestic rating

agencies mainly apply the former one which also has a large human impact factor, therefore the corporate rating results often do not objectively reflect the actual credit status of a company.

[0005] Existing quantitative methods for determining corporate credit ratings are also difficult to implement due to heavy manual maintenance of the corporate credit rating which creates delay between the actual credit quality change from the business and corporate credit rating update from rating agency which cannot reflect the credit quality change in real time.

[0006] It is further impossible to audit the rating outcome by an independent auditor due to the fact that the existing corporate rating method is mainly driven by the subject matter expert, and the final rating for each factor is hard to be audited by the auditor. Furthermore, an independent auditor cannot set up a unique standard to quantify the quality of corporate credit rating generated by different rating agencies.

[0007] Another problem is the lack of independence between credit rating agencies and the rated companies. According to the business model of a rating agency, a company which needs to get corporate credit rating for Initial Public Offerings (IPO) or bond issuing purpose, pays to the rating agency for the credit rating service. Since the demand of companies is to get a high credit rating for capital raising purposes and they are the party which pays fee to the rating agency, there is a high dependency between the credit rating agency and rated companies. This creates potential risk for the investors who rely on the corporate credit rating to make their investment decision including the wholesale investor (Bank, Pension) and the individual investor. This is also the root cause of 2008 GFC in the U.S.A.

[0008] The effectiveness of the existing rating methods in the non-developed country market is hard to guarantee as the current mainstream rating methodology is biased towards subjective judgment. Due to large interference of human factors, the fairness and objectivity of rating results cannot be guaranteed.

SUMMARY

[0009] The following summary is provided to facilitate an understanding of some of the innovative features unique to the disclosed embodiments and is not intended to be a full description. A full appreciation of the various aspects of the embodiments disclosed herein can be gained by taking into consideration the entire specification, claims, drawings, and abstract as a whole.

[0010] The present disclosure relates generally to a system, method and computer readable medium for determining a corporate credit rating. In particular, embodiments of the present disclosure are related to a computer-implemented semi-supervised machine learning system, method and computer readable medium for determining a corporate credit rating based on a multi-factor distance algorithm.

[0011] The system, method and computer readable media disclosed in the present application are advantageous in several ways including, but not limited to, providing a rating method for providing accurate and reliable ratings for listed companies based on distance definition, thereby preventing the biased expert judgment factors to impact the final corporate credit rating, using multiple evaluation factors (both quantitative and qualitative) to provide a quantified quantitative calculation of corporate credit rating, enabling access to clustering methods to get different industry groupings and successfully reducing the difference of financial indicators caused by industry heterogeneity, defining the 'good' and 'bad' clusters as the criterion in order to calculate the distance between the enterprises in different industry groups and the two clusters for measuring the performance of the enterprise, providing corporate credit rating from the weighted distance value, allowing a rating agency to provide up-to-date rating based on the credit quality change for a given company, solving excessive subjective problems caused by excessive artificial factors, providing a rating method driven by semi-supervised machine learning which is more transparent to the auditors and investors, providing auditable and comparable rating outcomes which allows independent auditors to provide assessment result for the credit rating to general investors.

[0012] In a first aspect of the present disclosure, a computer-implemented semi-supervised machine learning system for determining a corporate credit rating based on a multi-factor distance algorithm is disclosed. The system comprises: a data collection module for collecting at least one corporate data and a listed company information from at least one data provider; a data processing module for processing an enterprise data by selecting one or more financial and non-financial dimension indicators to reflect a capacity status of an enterprise; a model development module comprising the multi-factor distance algorithm for calculating a weighted distance and a rating result for the enterprise data; a Graphical User Interface (GUI) module to display the rating result on a GUI, wherein the rating result is based on the weighted distance; and a data storage module to store the corporate data, the enterprise data, the rating result.

[0013] According to an embodiment in conjunction to the first aspect of the present disclosure, wherein the data processing module classifies the listed company information based on at least one industry group using a k-means clustering algorithm by selecting at least one industry financial indicator.

[0014] According to an embodiment in conjunction to the first aspect of the present disclosure, wherein the data processing module defines a good sample clusters and a bad sample clusters of the listed company information based on consensus of a stock market.

[0015] According to an embodiment in conjunction to the first aspect of the present disclosure, wherein the data processing module performs standardization of the selected financial and non-financial indicators.

[0016] According to an embodiment in conjunction to the first aspect of the present disclosure, wherein the model developing module calculates a distance of enterprise data to a center of the good sample clusters and the bad sample clusters based on a standardized Euclidean distance.

[0017] According to an embodiment in conjunction to the first aspect of the present

disclosure, wherein the model developing module calculates a center vector of a cluster in the industry group.

[0018] According to an embodiment in conjunction to the first aspect of the present disclosure, wherein the model developing module calculates a distance of enterprise data to the center vector of the good sample clusters and the bad sample clusters based on standardized Euclidean distance.

[0019] According to an embodiment in conjunction to the first aspect of the present disclosure, wherein the rating result is based on a weighted distance value of enterprise data.

[0020] According to an embodiment in conjunction to the first aspect of the present disclosure, wherein the rating result is adjusted based on a public opinion information factor obtained from a public opinion adjustment module.

[0021] In a second aspect of the present disclosure, a computer-implemented semi-supervised machine learning method for determining a corporate credit rating based on a multi-factor distance algorithm is disclosed. The method comprises: collecting, by a data collection module, at least one corporate data and a listed company information from one or more data providers; processing, by a data processing module, an enterprise data by selecting one or more financial and non-financial indicator dimension to reflect a capacity status of an enterprise; calculating, by a model development module, a weighted distance and a rating result for the enterprise data; displaying, by a Graphical User Interface (GUI) module, the rating result on the GUI, wherein the rating result is based on the weighted distance; and storing, in a data storage module, the corporate data, the enterprise data and the rating result.

[0022] According to an embodiment in conjunction to the second aspect of the present disclosure, wherein processing of the enterprise data by the data processing module, comprising: classifying the listed company information based on at least one industry group using a k-means clustering algorithm by selecting at least one industry financial indicators; defining a good sample clusters and a bad sample clusters of the

listed company information based on consensus of a stock market; selecting at least one dimension indicator to reflect a capacity status of the enterprise; and performing standardization of the selected financial and non-financial indicators.

[0023] According to an embodiment in conjunction to the second aspect of the present disclosure, wherein calculating, by a model development module, a weighted distance and a rating result for the enterprise data comprising: calculating a center vector of a cluster in the industry group; calculating a weighted distance of enterprise data to the center vector of the good sample clusters and the bad sample clusters based on standardized Euclidean distance; and adjusting the rating result based on a public opinion information factor obtained from a public opinion adjustment module.

[0024] In a third aspect of the present disclosure, a computer readable medium storing a program comprising instructions that, when executed by a system, cause the system to perform the computer-implemented semi-supervised machine learning method for determining a corporate credit rating based on a multi-factor distance algorithm is disclosed.

BRIEF DESCRIPTION OF THE DRAWINGS

[0025] **FIG. 1** is a block diagram depicting an example system **100** for determining corporate credit rating, in accordance with the disclosed embodiment.

[0026] **FIG. 2** is a flow diagram that illustrates an example method **200** for determining corporate credit rating, in accordance with the disclosed embodiments.

[0027] **FIG. 3** is a flow diagram that illustrates an example data processing method **300** for determining corporate credit rating, in accordance with the disclosed embodiments.

[0028] **FIG. 4** is a diagrammatic representation of an example classification **400** that illustrates the clusters of classified industry groups, in accordance with the disclosed embodiments.

[0029] FIG. 5 is a flow diagram that illustrates an example method 500 for determining weighted distance and corporate credit rating, in accordance with the disclosed embodiments.

[0030] FIG. 6 is a diagrammatic representation of frequency distribution 600 of the corporate credit ratings, in accordance with the disclosed embodiments.

[0031] The summary above, as well as the following detailed description of illustrative embodiments, is better understood when read in conjunction with the appended drawings. For the purpose of illustrating the present disclosure, exemplary constructions of the disclosure are shown in the drawings. However, the disclosure is not limited to specific methods and instrumentalities disclosed herein. Moreover, those in the art will understand that the drawings are not to scale. Wherever possible, like elements have been indicated by identical numbers.

DETAILED DESCRIPTION OF ILLUSTRATIVE EMBODIMENTS

[0032] The particular configurations discussed in the following description are non-limiting examples that can be varied and are cited merely to illustrate at least one embodiment and are not intended to limit the scope thereof.

[0033] Corporate credit ratings are used in the finance industry as one of the deciding factors on whether a borrower does or does not receive a loan, and also helps investors in making investment related decisions.

[0034] The present invention generally relates to a framework that combines multiple evaluation factors (both quantitative and qualitative) and provides a quantified calculation of corporate credit rating. Commonly, the mainstream corporate credit rating methods determine the weights of risk factors through subjectivity, which are highly influenced by human interference. Unlike such methods, the present invention is mainly based on data quantified calculations to determine corporate credit rating. In the present invention, clustering method is used to obtain different industry groupings,

thereby reducing the difference of financial indicators caused by industry heterogeneity. Further, by defining “good” and “bad” clusters as the criterion, weighted distance between companies in different industry groups and the two clusters is calculated to measure the performance of a company. Finally, corporate credit rating is given based on the weighted distance value.

[0035] Various embodiments of the present invention relate to a system, method and computer readable medium for determining quantified corporate credit rating, which combines a small amount of labeled data with a large amount of unlabeled data and considers multiple evaluation factors (both quantitative and qualitative indicators) pertaining to a company to measure the weighted distance for determining corporate credit rating. Quantitative indicators, as the term denotes, may be understood as a quantity such as a pure number, an index, ratio or percentage, which give a very clear measure and are numerically comparable. Qualitative Indicators, on the other hand, are subjective as they depict the status of a company in more of the qualitative terms, for example positive and negative information about the company's public opinion. Quantitative indicators can include both financial and non-financial indicators, wherein financial indicators are the performance indicators that indicate the status of a company by its financial factors and economic terms. Financial indicators may include, but not limited to, registered capital, net profit, debt asset ratio, etc. Non-financial indicators are quantitative measures that cannot be expressed in numerical terms. Non-financial indicators may include, without limitation, staff number, listing age etc.

[0036] The system, method, and computer readable medium disclosed in the present invention combine semi-supervised learning with a multi-factor distance algorithm. Semi-supervised learning is an approach to machine learning that combines a small amount of labeled data with a large amount of unlabeled data during training. In semi-supervised learning, an algorithm learns from a dataset that includes both labeled and unlabeled data, usually mostly unlabeled. Generally, the corporate credit ratings do not have labeled data. Therefore, the disclosed system, method and computer readable medium in the present invention provide rules to label a small number of data samples, and expand to a large number of unlabeled data samples with the labeled

small number of samples. In the multi-factor distance algorithm, both financial and non-financial indicator with distance data of "good" sample clusters (referred as, good sample clusters) and "bad" sample clusters (referred as, bad sample clusters) are used.

[0037] FIG. 1 is a block diagram depicting an example system **100** for determining corporate credit rating, in accordance with the disclosed embodiment. The system **100** mainly comprises a data collection module **102**, a data processing module **104**, a model development module **106**, a public opinion adjustment module **112**, a rating result Graphical User Interface (GUI) module **108** and a data storage module **110**.

[0038] The data collection module **102** collects corporate data pertaining to at least one company from one or more data providers. Data providers can be understood as public or private agencies providing access to corporate data on various companies. An example of such a data provider is the National Enterprise Credit Information Publicity System which is a Chinese governmental credit information agency. Corporate data can be defined as any and all quantified data maintained by a company including, but not limited to, data related to its finances, taxes, employees, and so on.

[0039] Enterprise information datasets are obtained from the data providers for model training. The data collection module **102** extracts the enterprise data or information from the enterprise information datasets. The characteristic attributes of the enterprise data corresponding to each piece of information will constitute a characteristic attribute set. In some embodiments of the present invention, the data providers may also provide a listed company information. The enterprise data collected by the data collection module **102** is processed by the data processing module **104**.

[0040] The data processing module **104**, classifies or groups the listed company information into one or more industry groups based on at least one industry group by selecting at least one industry financial indicator. In some embodiments of the present invention, k-means clustering may be used for classification into industry groups. k-means clustering is a method that is used to partition n observations into k clusters in

which each observation belongs to the cluster with the nearest mean.

[0041] Then, the data processing module **104** defines a 'good' sample cluster and a 'bad' sample cluster of the enterprise data under each industry group based on a consensus of a stock market. For example, samples with 'blue chip' shares are classified as 'good' whereas the samples with 'ST' shares are classified as 'bad'. In the data processing module **104**, company dimension attributed by one or more financial and non-financial indicators is selected to evaluate the corporate credit rating of a company. The financial and non-financial indicators involved in the evaluation of corporate credit ratings are referred to as evaluation indicators. The enterprise data is prepared to be operated upon by the model development module **106** by performing min-max standardization on the evaluation indicators selected before. Min-max standardization is a feature scaling method that is used to normalize the range of independent variables or features of data. In data processing, min-max standardization is also known as data normalization and is generally performed during the data preprocessing step.

[0042] The model development module **106** applies the multi-factor distance algorithm on the processed enterprise data. Initially, the model development module **106** calculates center vector of each cluster in the industry groups formed in the data processing module **104**. Later, a weighted distance of the enterprise data to the center of the good sample clusters and the bad sample clusters is calculated based on standardized Euclidean distance. The model development module **106** then calculates corporate credit rating of a company based on the weighted distance. It may be noted here, that the corporate credit rating of the company has been calculated purely based on quantitative factors, that is, the processed enterprise data. The public opinion adjustment module **112** allows the corporate credit rating to be adjusted based on public opinion information which have been defined as qualitative factors. In some embodiments of the present invention, the public opinion information may be provided by one or more data providers. The GUI module **108** displays the corporate credit rating result on a GUI **109** of the system **100**. The data storage module **110** maintains a database which stores the corporate data, enterprise data, corporate credit rating information of all the companies and the related risk characteristic variable information.

[0043] **FIG. 2** is a flow diagram that illustrates an example method **200** for determining corporate credit rating, in accordance with the disclosed embodiments. As disclosed in **FIG. 2**, the example method **200** can be implemented, for example, at step **202** by collecting quantified corporate data pertaining to at least one company from one or more data providers using a data collection module. Data providers are the agencies which provide corporate data on various companies. The example method **200** can be implemented, as at step **204**, by processing the enterprise data by selecting one or more financial and non-financial indicator dimension to reflect a capacity status of a company using a data processing module. The processing of the quantified enterprise data includes, classifying the listed company information using a k-means clustering into one or more industry groups, defining 'good' and 'bad' sample clusters, selecting one or more financial and non-financial indicators (also referred to as evaluation indicators) related to a company dimension for the purpose of evaluating corporate credit rating of a company and then, finally standardizing the enterprise data by processing on evaluation indicators.

[0044] As disclosed in **FIG. 2**, the example method **200** can be implemented at step **206**, by calculating the weighted distance for the processed enterprise data and determining corporate credit rating of the company based on the weighted distance using a model development module. The corporate credit rating of the company obtained at step **206** is based purely on quantitative factors, that is, processed enterprise data. In some embodiments of the present invention, the quantitatively calculated corporate credit rating may be adjusted based on qualitative factors such as positive and negative information on the company's public opinion. Thereafter, the example method **200** can be implemented at step **208** by displaying the final corporate credit rating on a GUI, as at step **208**. Thereafter, as disclosed in **FIG. 2**, the example method **200** can be implemented at step **210** by storing the corporate data, the enterprise data and corporate credit ratings in the database maintained by the data storage module for future reference.

[0045] **FIG. 3** is a flow diagram that illustrates an example data processing method **300** for determining corporate credit rating, in accordance with the disclosed embodiments. As disclosed in **FIG. 3**, the data processing method **300** can be

implemented, for example, at step **302**, by classifying one or more industry groups by performing k-means clustering on the listed company information by selecting at least one industry financial indicator. In some embodiments of the present invention, the industry financial indicators are selected according to the industry classification standard of a stock market index for example, Shenwan index. In some embodiments of the present invention, the listed company information comprises a company name and its corresponding industry with one or more financial and non-financial indicator data.

[0046] In the present invention, k-means clustering has been used to classify ‘n’ sub-sectors into ‘k’ industry groups. Generally, for a given set of observations ($x_1, x_2, \dots, x_n$), where each observation is a d-dimensional real vector, k-means clustering aims at partitioning the n observations into k ($\leq n$) sets $S = \{S_1, S_2, \dots, S_k\}$, so as to minimize within cluster sum of squares $\|x - \mu\|^2$ (WCSS), that is the variance, by solving the minimization problem in Equation (1) quoted below:

$$\arg \min_S \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2 \quad \text{--- Equation (1)}$$

[0047] In Equation (1), “ μ_i ” refers to a mean of points in “ S_i ”. It may be noted that the minimization problem, that is, Equation (1) is solved to determine the variables according to actual application cases, and not fixed variables. .

[0048] As disclosed in **FIG. 3**, the example method **300** can be implemented, for example, at step **304**, by defining ‘good’ sample clusters and ‘bad’ sample clusters of the listed companies in each of the industry groups obtained at step **302**. In some implementations of the present invention, ‘good’ and ‘bad’ sample clusters of the listed company information may be defined under different industry groups based on the consensus from a stock market. ‘Good’ and ‘bad’ sample clusters refer to a sample of good and bad companies, respectively, that are classified according to their business definitions in the stock market. Examples of stock market may include, but not limited to, Chinese Stock Exchange, Shanghai Composite Index, and Shenzhen Stock Exchange. In some embodiments of the present invention, a sample company holding

‘blue chip’ shares can be classified as ‘good’, while a sample company holding ‘ST’ shares can be classified as ‘bad’.

[0049] As disclosed in **FIG. 3**, the example method **300**, can be implemented at step **306**, by selecting attributes of a dimension on which a company is likely to be provided a corporate credit rating. The attributes of the dimension include both financial and non-financial indicators. Since the corporate credit rating will be evaluated based on these selected financial and non-financial indicators, they are collectively referred to as evaluation indicators. In some implementations of the present invention, the evaluation indicators may be selected manually by an analyst or modeler based on their individual business experience. The following table discloses an example of a list of dimension indicators that may be used for selection of evaluation indicators.

Table 1

Dimension	Indicator	Dimension	Indicator
Business size	Registered capital	Compliance risk	Disclosure evaluation
	Staff number		Enforcement record
	Total asset		Abnormality record
	Listing age		Credit history
Debt-paying ability	Current ratio	Operational capability	Inventory turnover
	Quick ratio		Receivable turnover
	Debt asset ratio		Current assets turnover
	Total debt/EBIT		Total assets turnover
Profitability	Gross profit ratio	Development ability	Operating income growth rate
	Net profit margin		Owners' equity growth rate

Dimension	Indicator	Dimension	Indicator
	Return on Equity		Total Assets Growth Rate
	Return on Assets		Net profit growth rate

[0050] The example method **300**, can be implemented, at step **308**, by standardizing the enterprise data of the company by performing min-max standardization on the selected evaluation indicators in the enterprise data. Equation (2) below is the general formula of min-max standardization.

$$x^* = \frac{x - \min}{\max - \min} \quad \text{--- Equation (2)}$$

[0051] In some implementations of the present invention, at step **308**, data standardization is performed using the Equation (2) in order to convert the variable range of evaluation indicators to [0, 1].

[0052] **FIG. 4** is a diagrammatic representation of an example classification **400** that illustrates the clusters of classified industry groups, in accordance with the disclosed embodiments. Various companies in a listed company information can be clustered into industry groups. For example, according to the industry classification standard of a stock market, such as, Shenwan Index, a list of more than 3000 listed companies can be divided into clusters by k-means clustering. As a result of k-means clustering, the listed companies have been clustered into four industry groups, as illustrated by the classification **400**. The clusters depicted in the example classification **400** have been explained in **Table 2**.

Table 2

Cluster	No. of companies	Industry
0	1113	Asset-heavy, manufacturing-oriented industry
1	866	Asset-light (food,

Cluster	No. of companies	Industry
		garment), medicine, technology and entertainment industry
2	1203	Light manufacturing, chemical, electrical appliance industry
3	330	Construction, real estate, building decoration industry

[0053] FIG. 5 is a flow diagram that illustrates an example method 500 for calculating weighted distance and corporate credit rating, in accordance with the disclosed embodiments. As disclosed in FIG. 5, the example method 500 can be implemented, for example, at step 502, by calculating a center vector of each sample cluster in each of the industry groups. In some implementations of the present invention, the center of the 'good' and 'bad' sample clusters is defined as the point where the sum of the distances to the points in the sample cluster is the smallest. The sum of the distances is calculated using Equation (3) and the cluster center thus obtained is called as Fermat point. The specific variables used in Equation (3) can be referred from the Table 2 which discloses an example list of indicators.

$$\text{Dist}(x) = \sum_{i=1}^m \|x - p_i\| \quad \text{--- Equation (3)}$$

$$p(x_1, \dots, x_m), x_i \in (0, 1] \quad \text{--- Equation (4)}$$

[0054] The Fermat point is further optimized using Equation (4) and is referred to as the optimal point. The Equation (4) is an optimization problem which finds the optimal point that makes the sum of distances smallest. There are many methods available which can be used to solve the above optimization problem, for example, simulated annealing algorithm, newton iteration method and so on.

[0055] The example method **500** can be implemented, at step **504**, by calculating weighted distance of a sample enterprise data of a company whose corporate credit rating is being ranked. In some implementations of the present invention, a weighted distance of a sample enterprise data to a center of the good sample clusters or the bad sample clusters can be calculated based on standardized Euclidean distance.

[0056] At step **504**, standardized Euclidean distance is calculated using Equation (5) as the distance between the sample enterprise data of the company and the 'good' and 'bad' cluster center. The computed Euclidean distance is the measure of the difference between the sample enterprise data of the company and 'good' and 'bad' cluster. The equation (5) for calculating Euclidean distance is given below.

$$d = \sqrt{\sum_{i=1}^N \frac{(x_i - y_i)^2}{s_i^2}} \quad \text{--- Equation (5)}$$

[0057] As defined in the standardized Euclidean distance definition in Equation (5) above, Euclidean distance of each sample enterprise data is calculated for each of 'good' and 'bad' sample clusters in all the industry groups. Then, the weighted summation of calculated weighted distances of the samples enterprise data to the 'good' and 'bad' sample clusters is calculated.

[0058] Further, for a sample vector $\vec{x} = (x_1, \dots, x_n)$, the 'good' cluster center vector is taken as $\vec{y}_+ = (y_1, \dots, y_n)$, and the 'bad' cluster center vector is taken as $\vec{y}_- = (y_1, \dots, y_n)$. The distance of sample enterprise data to good or bad center is calculated based on standardized Euclidean distance in Equation (5), is denoted as d_- and d_+ . Then, rating index as defined in Equation (6) is calculated.

$$C = \frac{d_-}{d_- + d_+} \quad \text{--- Equation (6)}$$

At step **504**, the rating indices are sorted in descending order according to the value of the index in Equation (6). According to the present invention, the larger the value of a rating index, the higher is the corporate credit rating.

[0059] As disclosed in **FIG. 5**, the example method **500** can be implemented, as at

step **506**, by adjusting the corporate credit rating obtained in step **504** based on qualitative factors such as positive and negative information on the company's public opinion. In some implementations of the present invention, adjustment of the quantitatively evaluated corporate credit rating based on qualitative factors may be performed manually with reference to a rating agency. In some embodiments of the present invention, the quantitatively evaluated corporate credit rating is adjusted based on the number of records of the positive and negative information.

[0060] FIG. 6 is a diagrammatic representation of frequency distribution **600** of the corporate credit ratings, in accordance with the disclosed embodiments. The corporate credit rating results are obtained based on the weighted distance values and post adjustment based on qualitative factors. **Table 3** below discloses the example credit ratings that can be provided to a company based on a specific distribution.

Table 3

Rating	Grade	Risk
AAA	Investment	Lowest Risk
AA	Investment	Low Risk
A	Investment	Low Risk
BBB	Investment	Medium Risk
BB	Junk	High Risk
B	Junk	High Risk
CCC	Junk	Highest Risk
CC	Junk	Highest Risk
C	Junk	Highest Risk
D	Junk	In Default

[0061] The present invention solves the technical problem of subjective and non-quantitative evaluation of corporate credit rating methods used in prior arts, and provides a quantitative method for evaluating corporate credit rating based on

weighted distance definition in order to provide accurate and reliable ratings for a listed company. In this way, the present invention prevents biased expert judgment factors to impact the final corporate credit ratings.

[0062] The system, method and computer readable medium provided in the present invention cause the system **100** to perform the computer-implemented semi-supervised machine learning method for calculating a corporate credit rating of a company based on a multi-factor distance algorithm. It should be noted that the terms without limitation such as, “corporate”, “enterprise”, “company” or “corporation” are interchangeably used throughout the disclosure, either alone or along with other terms to imply the same or similar meaning.

[0063] The system, method and computer readable medium provided in the present invention use multiple evaluation factors, that is, both quantitative and qualitative, to provide a quantitative calculation of quantified corporate credit rating. By accessing various clustering methods to get different industry grouping, the present invention successfully reduces the difference of financial indicators caused by industry heterogeneity.

[0064] It will be appreciated that the present invention is more quantitative than the existing general corporate credit rating method (that is, expert method), which allows a rating agency, implementing the present invention, to provide up-to-date rating based on the credit quality change for a given company. Thus, the present invention effectively solves the excessive subjective problems caused by excessive artificial factors. Further, the semi-supervised machine learning algorithm of the present invention is more transparent to the auditors and investors. The model outcome of the present invention is auditable and comparable which allows even an independent auditor to provide assessment results for the corporate credit rating to general investors.

[0065] It will be further appreciated that variations of the above disclosed and other features and functions, or alternatives thereof, may be desirably combined into many other different systems or applications. Also, various presently unforeseen or

unanticipated alternatives, modifications, variations or improvements therein may be subsequently made by those skilled in the art which are also intended to be encompassed by the following claims.

[0066] Although embodiments of the current disclosure have been described comprehensively in considerable detail to cover the possible aspects, those skilled in the art would recognize that other versions of the disclosure are also possible.

CLAIMS

What is claimed is:

1. A computer-implemented semi-supervised machine learning system for determining a corporate credit rating based on a multi-factor distance algorithm, the system comprising:
 - a data collection module for collecting one or more corporate data and a listed company information from one or more data providers;
 - a data processing module for processing an enterprise data by selecting one or more financial and non-financial dimension indicators to reflect a capacity status of an enterprise;
 - a model development module comprising the multi-factor distance algorithm for calculating a weighted distance and a rating result for the enterprise data;
 - a Graphical User Interface (GUI) module to display the rating result on the GUI, wherein the rating result is based on the weighted distance; and
 - a data storage module to store the corporate data, the enterprise data, and the rating result.
2. The system of claim 1, wherein the data processing module classifies the listed company information based on one or more industry groups using a k-means clustering algorithm by selecting one or more industry financial indicators.
3. The system of claim 2, wherein the data processing module defines a good sample cluster and a bad sample cluster in the listed company information based on a consensus of a stock market.
4. The system of claim 3, wherein the data processing module performs a standardization of the selected one or more financial and non-financial

dimension indicators.

5. The system of claim 1, wherein the model developing module calculates a distance of the enterprise data to a center of the good sample cluster and the bad sample cluster based on a standardized Euclidean distance.
6. The system of claim 5, wherein the model developing module calculates a center vector of a cluster in the industry group.
7. The system of claim 6, wherein the model developing module calculates a distance of the enterprise data to the center vector of the good sample cluster and the bad sample cluster based on the standardized Euclidean distance.
8. The system of claim 1, wherein the rating result is based on a weighted distance value of the enterprise data.
9. The system of claim 8, wherein the rating result is adjusted based on a public opinion information factor obtained from a public opinion adjustment module.
10. A computer-implemented semi-supervised machine learning method for determining a corporate credit rating based on a multi-factor distance algorithm, the method comprising:
 - collecting, by a data collection module, one or more corporate data and a listed company information from one or more data providers;
 - processing, by a data processing module, an enterprise data by selecting one or more financial and non-financial dimension indicators to reflect a capacity status of an enterprise;
 - calculating, by a model development module comprising the multi-factor

distance algorithm, a weighted distance and a rating result for the enterprise data;

displaying, by a Graphical User Interface (GUI) module, the rating result on the GUI, wherein the rating result is based on the weighted distance; and

storing, in a data storage module, the corporate data, the enterprise data and the rating result.

11. The method of claim 10, wherein processing of the enterprise data by the data processing module, comprises:

classifying the listed company information based on at least one industry group using a k-means clustering algorithm by selecting at least one industry financial indicator;

defining a good sample cluster and a bad sample cluster of the listed company information based on a consensus of a stock market;

selecting at least one dimension indicator to reflect a capacity status of the enterprise; and

performing a standardization of the selected financial and non-financial dimension indicators.

12. The method of claim 10, wherein calculating the weighted distance and the rating result for the enterprise data comprises:

calculating a center vector of a cluster in the industry group;

calculating the weighted distance of the enterprise data to the center vector of the good sample cluster and the bad sample cluster based on a standardized Euclidean distance; and

adjusting the rating result based on a public opinion information factor obtained from a public opinion adjustment module.

13. A computer readable medium storing a program comprising instructions that, when executed by a system, cause the system to perform the computer-implemented semi-supervised machine learning method for

determining a corporate credit rating based on a multi-factor distance algorithm, of any of claims 10-12.