Applied Soft Computing 41 (2016) 305–316

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A patent quality analysis and classification system using self-organizing maps with support vector machine

Jheng-Long Wu[a,b], Pei-Chann Chang[a,c,*], Cheng-Chin Tsao[c], Chin-Yuan Fan[d]

[a] School of Software Engineering, Nanchang University, Nanchang, Jiangxi, China
[b] Institute of Information Science, Academia Sinica, Taipei, Taiwan
[c] Innovation Center for Big Data & Digital Convergence and Department of Information Management, Yuan Ze University, Taoyuan, Taiwan
[d] Science & Technology Policy Research and Information Center, National Applied Research Laboratories, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 1 July 2015
Received in revised form 6 August 2015
Accepted 13 January 2016
Available online 22 January 2016

Keywords:
Patent analysis
Patent quality
Data clustering
Patent quality classification
Machine learning

ABSTRACT

A plethora of patents are approved by the patent officers each year and current patent systems face a solemn quandary of evaluating these patents' qualities. Traditional researchers and analyzers have fixated on developing sundry patent quality indicators only, but these indicators do not have further prognosticating power on incipient patent applications or publications. Therefore, the data mining (DM) approaches are employed in this article to identify and to classify the new patent's quality in time. An automatic patent quality analysis and classification system, namely SOM-KPCA-SVM, is developed according to patent quality indicators and characteristics, respectively. First, the self-organizing map (SOM) approach is used to cluster patents published before into different quality groups according to the patent quality indicators and defines group quality type instead of via experts. The kernel principal component analysis (KPCA) approach is used to transform nonlinear feature space in order to improve classification performance. Finally, the support vector machine (SVM) is used to build up the patent quality classification model. The proposed SOM-KPCA-SVM is applied to classify patent quality automatically in patent data of the thin film solar cell. Experimental results show that our proposed system can capture the analysis effectively compared with traditional manpower approach.

1. Introduction

In recent years, more and more investing companies are eager to know which patent to invest on for future products, since hot product technology are playing an important role in making profit. However, the number of patents filed every year is incrementing at an expeditious speed. Thus the patents examined and approved in patent systems have a crucial role in the industry over different countries, and the research on patent's quality is getting more attention from the academic researchers and industrial practitioners. Patent quality analysis provides an expedient way through which companies determine whether or not to modify and continue manufacturing innovative products, but the question of how to evaluate and predict the quality or value of a new patent presents a new challenge to the researchers and industrial practitioners. Nevertheless, the current technology in patent quality evaluation and classification is still in its primitive stage, and the patents' qualities are still evaluated manually by the experts, leading to inaccuracies of all types.

Currently, there are various tools that are being utilized by organizations for analyzing patents. However, an important issue of patent analysis is patent quality analysis. The high-quality patent information can ensure success for business decision-making process or product development [1,2]. This study reviewed the patent analysis approaches that can understand patent status like patent quality, novelty, litigation, trends and so on [3]. However, traditional patent analysis requires spending much time, cost and manpower. The potential patents for high-quality determining approach need to have shortened analysis at times. In general, the analysis approaches are statistical analysis or indicators computation. Recently, the clustering method is widely applied to cluster patents according to patent characteristics for patent trend [4]. The methods with statistical analysis can help analysts to understand patent situation or trend of this time, but if we want to know the potential quality of a newly applied patent, it does not provide effective rules or solutions to determination. The future patent

[*] Corresponding author at: Innovation Center for Big Data & Digital Convergence and Department of Information Management, Yuan Ze University, Taoyuan, Taiwan. Tel.: +886 936101320; fax: +886 34638884.
E-mail address: iepchang@saturn.yzu.edu.tw (P.-C. Chang).

http://dx.doi.org/10.1016/j.asoc.2016.01.020
1568-4946/© 2016 Elsevier B.V. All rights reserved.

evaluation is a key issue when a new patent is applied or published because patent has been producing impact on the industry according to the past industrial development such as patent litigation, specifically high-tech or information.

The patent officers approve a large amount of patents each year and current patent systems face a serious problem of evaluating these patents' qualities. Traditional researchers and analyzers have focused on developing various patent quality indicators. The patent indicators are collected from patent corpuses, including the number of patent citations and the number of International Patent Classifications (IPC). The primary patent quality indicators [5–7] are related to investment, maintenance, and litigation, which form a basis for assessing patent. But, these indicators do not have further predicting power on a new patent application or publication. Therefore, the data mining (DM) approaches are employed in this article to identify and classify the new patent's quality. Investors from venture capital companies can derive these patents' qualities in time when making decisions regarding the development of new innovative products and discovering the trend of state-of-the-art technology. Thus, an automatic classification system to analyze and forecast the patent quality is needed in order to quickly respond important or emergency situations.

In this study, we propose an automatic patent quality classification system named SOM-KPCA-SVM which combines three DM methods including self-organizing maps (SOM), kernel principal component analysis (KPCA) and support vector machine (SVM). The SOM is a two-dimensional (or multidimensional) network structure for multiple variables mapping and cluster sample to several groups [8]. Therefore, SOM in this article is used to cluster patents into several quality groups according to patent quality indicators. We will summarize these quality indicators in order to delimit different quality for each group. The result of quality analysis on patent data extends to a classification problem in order to early identify valuable patent as well as patent quality forecasting. In addition, the KPCA is based on kernel mapping with principal component analysis (PCA). KPCA is used to transform original feature space into a new nonlinear feature space through nonlinear kernel mapping and relationships among new feature that are independent variables [9]. Thus, we will apply KPCA to extract key characteristics of a patent from the patent document. In the classification problems, the SVM classification approach is a powerful tool for solving many kinds of problems such as stock trends [10] and patent classification [11]. In this article, the SVM is used to build patent quality classification model and it can automatically determine the patent quality and hence there is no need to hire an expert to rank or define patent quality. Therefore, a SOM-KPCA-SVM system can automatically analyze patent quality based on past patent applications and forecast an unknown patent's quality, better enabling engineers and product designers to forecast a patent's potential for product development or innovation.

The inventors, attorneys, examiners, governments and companies need to reach a consensus on the quality of patents. Though the value can be estimated manually or by experts studying about the actual quality decisions, this is slow and expensive. In this study, we introduce an automatic analysis and classification system of patent quality named SOM-KPCA-SVM, which represents the quality in which the application will be classified. The main contributions of this work are summarized as follows:

- The quality type is identified based on the clustering approach by self-organizing maps.
- New feature sets of nonlinear space are transformed by KPCA for improving the quality classification of patent applications.
- Quality classification system to classify the patent quality is obtained by using a support vector machine.
- The classification effectiveness of the model is shown by an evaluation using more than 18,000 applications around the world related to thin film solar cell.
- A new patent analysis system allows users to automatically analyze the quality of patent applications.
- A group of intellectual property experts who work with solar cell hi-tech companies obtained analysis results by classifying patent applications on the quality classification system.

The overall structure of this article is as follows. Literature review is given in Section 2. The proposed patent quality analysis and classification system by SOM, KPCA and SVM is introduced in Section 3. Following the applications and results, discussions of the proposed system are shown in Section 4. Conclusions and future research directions are finally presented in Section 5.

2. Literature reviews

In this section, first we summarize the studies related to the patent analysis. Second, we focus the patent quality indicators that are used in this study for calculating patent quality. Third, literatures in SOM are reviewed for patent clustering analysis. Fourth, the concepts of feature extraction of KPCA are reviewed for capturing key features. Finally, we review literature in SVM for developing the patent classification system.

2.1. Patent analysis

The patent analysis is a set of techniques and visual tools that analyze the trend and patterns of technology innovation in a specific domain based on statistics of patents. The analysis objects on patent include [12–14]: (1) patent count analysis, it is counting the quantity of patents, including the technology life cycle chart and the patent quantity comparison chart; (2) country analysis, it is comparing the patents of various countries in a specific technology domain; (3) assignee analysis, it is comparing detailed data on R&D, citation ratio, cross-citation, event charts, ranking chart, and competitors; (4) citation rate analysis, it is comparing the number of citation made by other patents during its valid period; and (5) International Patent Classification (IPC) analysis, it is including IPC patent activity chart and number of IPC patent companies. There are various tools utilized by organizations for analyzing patents. These tools are capable of performing a wide range of tasks, such as forecasting future technological trends, detecting patent infringement and determining patent quality. Moreover, patent analysis tools can free patent experts from the laborious tasks of analyzing the patent documents manually and determining the quality of patents. The tools assist organizations in making decisions of whether or not to invest in manufacturing of the new products by analyzing the quality of the filed patents [2]. This eventually may result in imprecise recommendation of patents. Grimaldi et al. [15] proposed a patent portfolio value analysis that uses the leverage page patent information for strategic technology planning. They used five directions of patent quality such as claim, citation, market coverage, strategic relevance and economic relevance to analyze patent portfolio value. Another study [16] used renewal data to estimate the value of US patents. In their analysis results, they mentioned that the ratio of US patent value to R&D is only about 3% but these had a high value in terms of litigation.

2.2. Patent quality indicator

Usually the patent quality indicators are related to investment, maintenance, litigation and patent claims. The quality is used to evaluate the potential patents for a business or government policy [17,18]. For example, one kind of indicator of patent quality is legal status (LS), which means the technology's potential. In this movement, the competitors may litigate a claim in this patent. Therefore, the quality indicators can respond to the future value for business intelligence. Archontopoulos [19] mentioned that the legal status and search tools on the Internet are so sensitive that emphasis given to the issues related to the date of availability to the public has a wide Internet disclosure. Simmons and Spahl [20] mentioned that many changes in the legal status of US patents are hidden away in files at the United States Patent and Trademark Office USPTO and at the Courts of Appeal. Patent applications have "first to invent" concepts in US patent law; therefore, legal status of previous patent has been affecting other patent judgments. Hirschey and Richardson [21] mentioned that the scientific measures of patent quality have the potential to offer managers useful guidance concerning the quantity and quality of inventive output and the effectiveness of patent investments. The quality indicators are current impact index, total patent, science linkage and technology cycle time for patent quality valuation analysis for innovation level in a Chinese patent [22]. They merged the patent data with widely used industrial survey data by firm names which made it possible to investigate the relationship between financial performance and innovation activities. Chen and Chang studied [23] the relationship between a firm's patent quality and its market value. The relative patent position, revealed technology advantage, Herfindahl– Hirschman Index of patents, and patent citations are used to analyze their patent data. In their results, the pharmaceutical companies with high patent counts had higher marker value than those with low patent counts. Frietsch et al. [24] mention that the values of patents were estimated by export volume. Their results show that exports are a very useful way of placing a valuation on patents according to the patent applications, forward citations, in particular analysis. Another study [25] used the citations, family size, and opposition to estimates value of patent right. In their analysis result they mentioned that the number of citations a patent receives are positively related to its value. References to the non-patent literature are informative about the value of pharmaceutical and chemical patents, but not in other technical fields.

2.3. Self-organizing maps

The SOM is a two-layer neural network that maps multidimensional data on to a two-dimensional topological grid. The data are grouped according to the similarities and patterns found in the data set, using some form of distance measure approach such as the Euclidean distance. The results are displayed as nodes on the map, which can be divided into different clusters based upon the distances between the clusters. Since the SOM is unsupervised, no target outcomes are provided, and the SOM is allowed to freely organize itself. So the SOM is an ideal tool for exploratory data analysis. Segev and Kantola [8] used SOM to identify patent trends and analyzed patent knowledge to identify research trends. They tested on patents from the USPTO and provided an overview of the directions of the trends. Another one of SOM algorithm is called evolving self-organizing map (ESOM), it is an evolving network structure and has fast learning rate on-line. Their result shows that ESOM achieved better or comparable performance with a much shorter learning process [26]. Juntunen et al. [27] used the SOM combined with K-mean clustering to a model of water quality in a treatment process. They provided approach offers a straightforward clustering method for assessing the essential characteristics of the process. Their results show that modelling of water quality in treatment process has clearly demonstrated some challenges. Bouhouche et al. [28] combined PCA and SOM for condition monitoring in pickling process. In their research, the PCA method uses indexes to classify process variability and the SOM is used to replace the conventional indexes. Their result shows that the combined approach remains important comparatively to PCA but not more than SOM.

2.4. Kernel principal component analysis

Principal component analysis (PCA) is very useful to extract nonlinear features for many research applications. The PCA is widely used in many fields of research. Trappey et al. [1] tried to identify the key impact factors using PCA and the authors selected a lot of variables as indicated by the first five components. The PCA was also used to estimate global competition using patent statistics and analysis [29]. In the further research, the kernel principal component analysis is an extension from PCA that uses the kernel mapping process before the Eigen-problem. PCA is mathematically defined. The KPCA is used for critical feature extraction in stock trading model which they used Gaussian and Euclidean kernel function to feature space transformation [6]. In their experimental results, the KPCA has captured best performance compared to PCA, ICA and so on. Shao et al. [30] used the PCA and KPCA for fault feature extraction. Their study compared and analyzed the extraction effect for primary feature data and feature dimensionality reduction. The effect of feature extraction is KPCA approach which is caused by the kernel function.

2.5. Support vector machine

Support vector machine is a machine learning algorithm which is widely used for classification and regression problems. Vapnik [31], and Cortes and Vapink [32] developed the SVM that can solve multidimensional classification problems. This method aims to develop an optimal hyper-plane as a decision function using the maximum margin hyper-plane between class vectors on both sides of the hyper-plane in binary classification problem. SVM map inputs vectors into the high dimensional feature space via the linear or non-linear transformation. An effective decision hyper-plane is developed to distinguish the correct training data and it can determine class in testing data according to known hyper-plane. An approach is proposed integrated with a hybrid genetic-based support vector machine (HGA-SVM) model for developing a patent classification system [33]. But the target of classification must be definition by experts' analysis and knowledge. The authors claim that they use these models in real-world cases of patent classification rather than using only for International Patent Classification (IPC). The study integrated the honeybee mating optimization algorithm with SVM (HBMOSVM) for patent document categorization. In their results they showed that the HBMOSVM could result in a better patent documentation accuracy and better F-measure performance as an evaluation index than GASVM model in patents document categorization [34]. On the other hand, SVM classifier is used to classify the topics of patents and their results are based on the text data to build classifier and data from USPTO [35].

3. Proposed method

This study proposed an automatic patent quality classification system that integrated a system combining three approaches including self-organizing maps, kernel principal component analysis and support vector machine, namely SOM-KPCA-SVM. This quality classification system has two stages to implement: stage one is patent analysis and quality definition, and stage two is a patent quality classification model building as shown in Fig. 1. In stage one, we collect the patent data related specific industry from patent database and the SOM approach will be used to cluster patents into several quality groups with summarized quality indicators in order to identify each group quality class. In stage two, first, the KPCA approach will be used to extract key characteristics

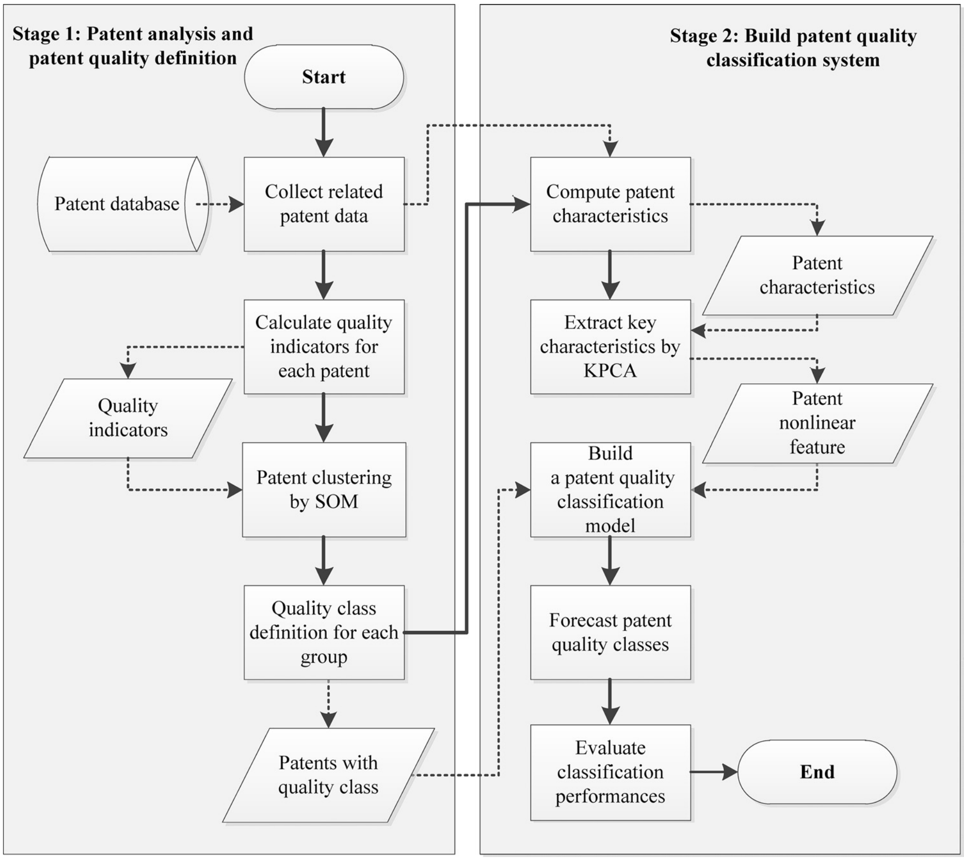
Stage 1: Patent analysis and patent quality definition
Start
Patent database
Collect related patent data
Calculate quality indicators for each patent
Quality indicators
Patent clustering by SOM
Quality class definition for each group
Patents with quality class
Stage 2: Build patent quality classification system
Compute patent characteristics
Patent characteristics
Extract key characteristics by KPCA
Patent nonlinear feature
Build a patent quality classification model
Forecast patent quality classes
Evaluate classification performances
End

Fig. 1. The framework of SOM-KPCA-SVM patent quality classification system.

from linear to nonlinear feature space, in which the original feature is by patent document, and then, the SVM classifier will be used to build patent classification model using nonlinear feature in order to improve classification performance. Finally, we forecast quality for each patent and evaluate this patent quality classification system performance by our proposed system. Our proposed system is developed as follows:

3.1. *Stage 1: Patent analysis and patent quality definition*

In this stage, we must know the current trend of patent quality from past years. Therefore, we need to analyze the phenomenon of patent quality based on patent quality indicators and cluster them into different kinds of quality groups. Next, we will define which patent has which quality type according to SOM clustering analysis. The details of quality analysis and patent quality definition are following:

(1) Patent data collection, quality indicators calculation and patent characteristics computation

This patent data in a specific industry will be collected from patent database such as WebPat, PatBase, Google Patent Search, Thomson Innovation, patent office of country and so on. These patent offices can provide all the patent documents. The patent service organization can provide other information such as legal status. Therefore, our patent data include characteristics of the patent document and quality indicators from additional information. The quality indicators and characteristics will be normalized using the Min-Max method.

(2) Patent quality analysis using SOM approach based on quality indicators

The patent clustering is adopted in the SOM approach, to identify patent quality and explore hidden patterns among these patents. According to this algorithm, these patents will be split to several groups (clusters) in order to cluster a similar quality patents based on quality indicators. Therefore, we use SOM clustering method based on quality indicators to analyze and cluster patents. The analysis process continues until all input vectors are processed. Convergence criterion utilized here is in terms of epochs, which defines how many times all input vectors should be fed to the SOM for analysis. The details of the SOM algorithm in this study are the following:

- Step 1: Set-up the parameters in the SOM network such as the number of clusters and epoch.
- Step 2: Initialize each neuron weight $w_i = [w_{i1}, w_{i2}, .., w_{ij}]^T \in \Re^j$. In this study, neuron weights are initialized by drawing random samples from input dataset.
- Step 3: Present an input pattern $x = [x_1, x_2, .., x_j]^T \in \Re^j$. In this case, the input pattern is a series of variables representing current patent status. Calculate the distance between pattern x, and each neuron weight w_i, and therefore, identify the winning neuron or best matching unit c, as in Eq. (1).

$$\|x - w_c\| = \min_i \left\{d_i\right\}, \tag{1}$$

$$d_i = \sqrt{\sum_j (x_j - w_{ij})^2} \tag{2}$$

- Step 4: Adjust the weight of winning neuron c and all neighbor units, the adjustment as in Eq. (3).

$$w_i(t+1) = w_i(t) + h_{ci}(t)[x(t) - w_i(t)], \tag{3}$$

where *i* is the index of the neighbor neuron and *t* is an integer, the discrete time coordinate. The neighborhood kernel $h_{ci}(t)$ is a function of time and the distance between neighbor neuron *i* and winning neuron $h_{ci}(t)$ as in Eq. (4). It defines the region of influence that the input pattern has on the SOM and consists the neighborhood function $h(\|\cdot\|, t)$ and the learning rate $\alpha(t)$.

$$h_{ci}(t) = h(\left\|r_c - r_i\right\|, t)\alpha(t), \tag{4}$$

where *r* is the location of the neuron. In this work we used Gaussian Neighborhood Function. The learning rate function $\alpha(t)$ is a decreasing function of time. The final form of the neighborhood kernel with Gaussian function is given as follows.

$$h_{ci}(t) = \exp\left(\frac{\left\|r_c - r_i\right\|}{2\sigma^2(t)}\right)\alpha(t), \tag{5}$$

- Step 5: repeat steps 3 and 4 until the convergence criterion is satisfied. Average value for each variable of each clustered group was calculated after the patent cases were clustered, and the average value for each variable of each group would be the basis when finding the most matching group for the new case. After the set of patent data has been processed by SOM, a new case can be categorized into a pre-defined group.

(3) Patent quality definition for groups

The SOM is used to cluster patents into several groups according to patent quality indicators. Each group has a specific difference compared to the other. Therefore, we define different quality type for each group such as high, middle and low in three clusters. The average quality on each group Quality (group_g) based on normalized quality indicators is calculated as in Eq. (6).

$$\text{Quality}(\text{group}_g) = \frac{1}{m \times n}\sum_{i=1, i \in \text{group}_g}^{m}\sum_{j=1}^{n} q_{ij}, \tag{6}$$

where q_{ij} denotes value of quality of the *j*th quality indicator of *i*th patent in gth group. The *m* and *n* denote the number of patent in a group and the number of quality indicators, respectively. We can sort quality value of patent groups according to total quality value of groups, for example, if the number of groups is 3, the maximal quality value of patent group defines the highest quality, second high value defines middle quality and the minimal value is defining lowest quality. Therefore, each group has a quality degree for further classification building stage.

3.2. Stage 2: Build patent quality classification system

These quality classes are defined according to the SOM quality analysis based on the quality indicators. In this stage, the task of patent quality prediction is predicting quality potentiality on new patent application. There are two steps in extracting key characteristics of a patent document and building an SVM-based quality classification system according to patent characteristics. However, the quality indicators cannot be used directly in variables of classification system because the quality indicators are calculated at patent publication afterwards. The new patent may only include few or none of quality indicators. Normally, new patent application or just published has some potential but these indicators does not appear at this time. Therefore, we only use the characteristics from patent document such as count of claim, priority date and so on.

(1) Key patent characteristic extraction by KPCA

All patents will separate into two datasets, which are training and testing data. The characteristics of a patent document in training dataset are used to compute mean centering for kernel, eigenvalues and eigenvectors. In order to compute dot products of the form $(\Phi(x_i), \Phi(x_j))$, we use kernel representation of the form as in Eq. (7).

$$K(x_i, x_j) = (\Phi(x_i), \Phi(x_j)) \tag{7}$$

where x_i and x_j are vectors in the input space and allows us to compute the value of the dot product in a possibly high-dimensional feature space *F* without having to carry out the map Φ. A number of kernel functions exist and have been chosen before we apply the algorithm. The representative Gaussian kernel as in Eq. (8) and Euclidean kernel as in Eq. (9) in which two functions as *K* are described:

- The Gaussian or radial basis function kernel is given by:

$$K(x_i, x_j) = \exp\left(-\frac{\left\|x_i - x_j\right\|}{2\sigma^2}\right) \tag{8}$$

- The polynomial kernel of degree *d* is given by:

$$K(x_i, x_j) = (x \cdot y + c)^d \tag{9}$$

where $c \geq 0$ is a free parameter trading off the influence of higher-order versus lower-order terms in the polynomial. Given a set of *n*-dimensional normalized patent indices $x_k \in R^N$ of training data, we compute the kernel matrix $K \in R^{N\times N}$ from two kernels in Eq. (8) or Eq. (9),

$$K_{ij} = (\Phi(x_i), \Phi(x_j)) = [k(x_i, x_j)] \tag{10}$$

Carry out mean centering K' in the feature space *K* for $\mu = \sum_{k-1}^{N}\tilde{\Phi}(x_k) = 0$,

$$K' = K - C \times K - K \times C + C \times K \times C \tag{11}$$

where

$$C = \frac{1}{N}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \in R^{N\times N}$$

The *K* is by $K_{ij} = (\Phi(x_i), \Phi(x_j))$ and using it to solve the eigenvalue problem $N\lambda\alpha = K\alpha$. For non-zero eigenvalues, performing PCA in the feature space *F* is equal to resolving the Eigenproblem $N\lambda\alpha = K\alpha$. The patent data of identified quality will split into two datasets because of forecasting problem and needs to train a forecasting model to later forecast future by its model. The new feature of training data Training(tr^k) formed by the mean centered $\tilde{\Phi}(x)$ and α_i^k. For patent variables *x* in the training period, we extract a nonlinear component via Eq. (12)

$$\begin{aligned}\text{Training}(tr^k) &= \sum_{i=1}^{N}\alpha_i^k(\tilde{\Phi}(x_i), \tilde{\Phi}(x)) \\ &= \sum_{i=1}^{N}\alpha_i^k K(x_i, x)\end{aligned} \tag{12}$$

where $\tilde{\Phi}$ is the mean centered, α is the eigenvalue and x_i denotes the normalized variables in training data.

(2) Build a patent quality classification model by SVM

The input vector of SVM training use new nonlinear feature space Training(tr^k) by KPCA and the output vector of quality classes is given by SOM with quality indicators. The learning algorithm for a non-linear classifier SVM follows the design of an optimal separating hyper-plane in a feature space. The procedure is associated with hard and soft margin classifier SVMs. The dual Lagrange in the z-space as in Eq. (13)

$$L_d(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}y_iy_j\alpha_i\alpha_j z_i^T z_j \tag{13}$$

Further, by using the chosen kernels, we maximize the Lagrange as in Eq. (14).

$$\begin{aligned}&\text{Maximize}:\quad L_d(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}y_iy_j\alpha_i\alpha_j U\left(x_ix_j\right)\\&\text{Subject to}:\quad \alpha_i\geq 0,\quad i=1,\ldots,1 \qquad ,\\&\qquad\qquad\sum_{i=1}^{l}\alpha_i y_i=0\end{aligned} \tag{14}$$

where the U denotes the popular kernel functions in machine learning theories. The kernel also defined in Eq. (8) and (9).

Note that the constraints must be revised for use in a non-linear soft margin classifier SVM. The only difference between these constraints and those of the separable non-linear classifier are with regard to the upper bound C on the Lagrange multiple α_i. Consequently, the constraints of the optimization problem take the following form

$$\begin{aligned}&\text{Subject to}:\quad C\geq\alpha_i\geq 0\\&\qquad\qquad\sum_{i=1}^{l}\alpha_i y_i=0\ ,\end{aligned} \tag{15}$$

In this way, the influence of the training data point will be limited and will remain on the wrong side of a separating non-linear hyper-surface. The decision hyper-surface $d(x)$ and the indicator function, which were determined by the nonlinear SVM classifier, are as in Eq. (16) and (17).

$$d(x)=\sum_{i=1}^{l}y_i\alpha_i U(x_i,x_j)+b \tag{16}$$

$$i_F(x)=\text{sign}(d(x))=\text{sign}\left(\sum_{i=1}^{l}y_i\alpha_i U(x_i,x_j)+b\right) \tag{17}$$

Depending upon the chosen kernel, the bias term b may be an implicit part of the kernel function. For example, b is not required when Gaussian RBF are used as kernels. When d is included within other kernel functions, the non-linear SVM classifier is as in Eq. (18).

$$i_F(x)=\text{sign}(d(x))=\text{sign}\left(\sum_{s=1}^{\text{number of SVs}}y_s\alpha_s U(x,x_s)\right) \tag{18}$$

(3) Forecast patent quality for new sample of testing data

The trained model of SVM classifier will be used to classify the new sample of testing data Testing(ts^k) for patent quality classification. Thus, the purpose of this study is to build the patent quality classification model for evaluating the potential of patents. The testing data are unknown data as well as the future data set. The testing data will transform to non-linear feature space by KPCA transformation. We cannot use the mean centering of testing data to calculate new testing data in order to avoid the misuse of unknown data. We use $\tilde{\Phi}(x)$ and α_i^k from the training data to transform and extract a non-linear feature space on testing data. The Testing(ts^k) extraction is shown in Eq. (19).

$$\begin{aligned}\text{Testing}(ts^k)&=\sum_{i=1}^{N}\alpha_i^k(\tilde{\Phi}(y_i),\tilde{\Phi}(x))\\&=\sum_{i=1}^{N}\alpha_i^k\tilde{K}(y_i,x)\end{aligned} \tag{19}$$

Table 1
The confusion matrix for each patent quality class.

Class *i*		Actual judgment	
		True	False
Classifier judgment	True	TP_i	FP_i
	False	FN_i	TN_i

where y denotes the normalized variables $y_k\in R^m$ in testing data. The new non-linear feature space of testing data is obtained and used to evaluate performance of our proposed SOM-KPCA-SVM patent quality classification.

(4) Evaluation of patent quality classification performance

The general classification evaluation method works with three evaluations: Accuracy (ACC), Precision (PRS), and Recall (REC). Table 1 shows that the confusion matrix is for each quality class *i*, the TP_i denotes the correct classification into quality class *i*, the *TN* denotes the correct classification into non-quality class *i*, the FN_i denotes incorrect classification into quality class *i*, the FP_i denotes incorrect classification into non-quality class *i*.

We will use the three performance indicators to evaluate our proposed patent quality classification for all tests and the performance indicators of accuracy (ACC), precision (PRS) and recall (REC) performance formulas as in Eqs. (20)–(22):

$$\text{ACC}_i=\frac{\text{TP}_i}{\text{TP}_i+\text{FP}_i+\text{TN}_i+\text{TN}_i} \tag{20}$$

$$\text{PRS}_i=\frac{\text{TP}_i}{\text{TP}_i+\text{FP}_i} \tag{21}$$

$$\text{REC}_i=\frac{\text{TP}_i}{\text{TP}_i+\text{FN}_i}. \tag{22}$$

4. Experimental results

In this section, we have designed a series of testing for evaluating our proposed methodology as SOM-KPCA-SVM. There are three parameters for experiments, first one, the scale of data on time has three different period datasets which are five years, ten years and forty years; second one, the amount of quality groups has three direction which are three quality groups, five quality groups and seven quality groups; finally, the number of feature extraction has four percentages which are 25%, 50%, 75% and 100%.

4.1. Patent dataset and statistical analysis

In this patent data, we collected 18,747 patents in last 40 years from Thomson Innovation[1] database. These patents are related to "Thin film solar cell" technical field and search on title, abstract, claim and description. The Fig. 2 statistics shows a yearly publication patent related to a thin film solar cell from 1974 to 2013.

[1] The Thomson Innovation is proved the fully patent data from around the world, http://info.thomsoninnovation.com/.

Table 2
Patent quality indicators for patent analysis.

No.	Name	Explanation
PQI1	INPADOC legal status[a]	The legal status of each patent is quantified by the number of applications. The information of legal status is maintained by INPADOC
PQI2	INPADOC patent family[a]	Patent family is referred to as an "extended" family since it includes all family members sharing at least one priority number
PQI3	DWPI patent family country[b]	This patent family is based on DWPI patent family calculation, which means counting the number of countries
PQI4	DWPI patent family[b]	Patent family is referred to only same priority number and same technology claim family since it includes all family members that share one priority number
PQI5	Patent backward citation	This citation means the number of patents who cited this patent
PQI6	Non-patent backward citation	This citation means the number of non-patents who cited this patent

[a] INPADOC is International Patent Documentation. The database is produced and maintained by the European Patent Office (EPO) and contains patent families and legal status information.
[b] The database is produced by information provider Thomson Reuters.

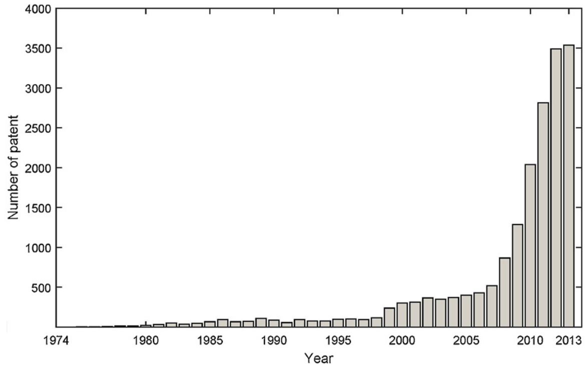

Fig. 2. Yearly published patents of a thin film solar cell.

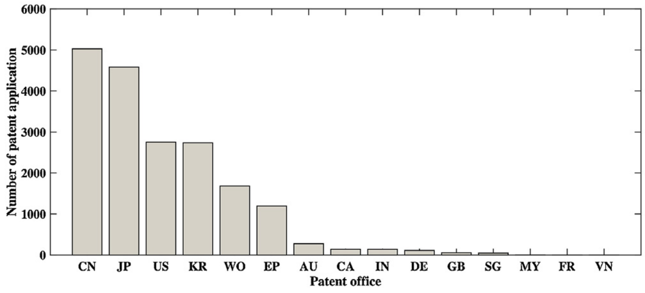

Fig. 3. Scale of patent publication on 15 patent offices (1974–2013).

The patent application in a thin film solar cell is rapidly growing up in recent years. In this study, we want to analyze the trends of different time scales, so there are three datasets:

- **5YD**: there have 13,164 patent applications in last 5 years (2009–2013).
- **10YD**: there have 15,751 patent applications in last 10 years (2004–2013).
- **40YD**: there have 18,747 patent applications in last 40 years (1974–2013).

At the end of the year 2013, there were total five patent offices including Australia (AU), Canada (CA), China (CN), Deutsche (DE), European (EP), France (FR), Britain (GB), India (IN), Japan (JP), Korea (KR), Malaysian (MY), Singapore (SG), United States (US), Vietnam (VN) and World patent (WO). There are six patent offices having most of the patents including CN, EP, JP, KR, US, and WO patent offices. The statistics of the result is shown in Fig. 3. The six patent offices have 17,971 patents which is 95.86% patent among last 40 years.

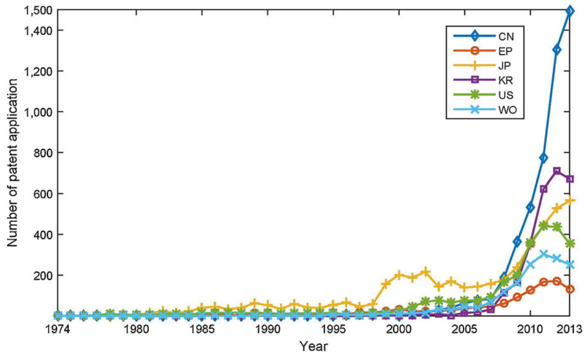

Fig. 4. The trend of published patents on major six patent offices.

In addition, the trends of published patents of six patent offices among 40 years are shown in Fig. 4. The publication trends of two patent offices are increasing and four patent offices are decreasing in 2013. However, the trends of all the patent offices are increasing in recent years and only JP patent office has a sharp increase from 1999 to 2007. In the year 2008, the number of patent application of CN office is more than that in JP office, and the patent applications in other patent offices are quickly increasing.

The variables of patent data are of two types, including patent quality indicators (PQI), which are calculated from the additional information, and patent document characteristics (PDC) of patent document. The six PQIs are selected according to references and experts, as shown in Table 2. They are used as input variables to SOM to cluster the patent into different quality groups. The quality indicators of each patent are from patent database of Thomson Reuters. As for eight PDCs shown in Table 3, document characteristics data of each patent are used as inputs to SVM to train and to build up a quality classification system. All PDCs are based on counting method in scalar values.

4.2. Result on patent quality analysis

In this patent analysis, the result of quality analysis of past patent development is obtained by SOM clustering computation. To decide the right number of groups, the rule of thumb is that the smaller the groups are, the larger the quality differences are. Of course, the number of patents is also an important factor in deciding number of groups. To properly cluster the patent data, in this study, we designed different quality groups, i.e., three, five and seven. In addition, we will look into the patents within each group and check with the quality indicators of each patent to ensure the consistency of the clustering. There are three different amounts of clustering such as 3 quality groups (3QG), 5 quality groups (5QG) and 7

Table 3
The patent document characteristics for the quality classification system.

No.	Name	Explanation
PDC1	Assignee	How many number of assignee of this patent
PDC2	Patent forward citation	The citation means the number of previous patents this patent cited
PDC3	Claims	How many technical areas this patent protected
PDC4	IPCs	The IPCs means the number of class in this patent
PDC5	Inventors	The inventor count number of assignee in a patent document
PDC6	Prosecution period	The prosecution period means period from the applied date to approval date
PDC7	Priority country	The priority is the number of priority countries advocated in this patent
PDC8	Priority period	The period means the period from the priority date to approval date

quality groups (7QG). Table 4 shows that the quality summary on three different amounts of quality group with three different datasets related to the thin film solar cell patent. The analysis result of three quality groups with 5YD dataset has the minimum quality score 0.0776, as a result, it is lower quality group (G1); the middle quality score 0.5419 as a result, it is delimited middle quality group (G2); and the maximum quality score 1.1213, as a result, it is high-quality group (G3). Therefore, all G1 groups are low quality patent groups with the lowest average quality in different time period datasets. All G3 in 3QG, G5 in 5QG and G7 in 7QG are the highest patent's quality group, respectively.

Figs. 5–7 show the distribution of patent applications in 3QG, 5QG and 7QG, respectively. The results show that all G1 with the lowest quality has the largest number of patent applications in all three different database. The highest quality groups such as G3 in 3QG, G5 in 5QG and G7 in 7QG in which only the G7 and G 5 have less number of patent applications. The reason is that usually higher quality patents are only relatively few, especially when the number of groups is increasing.

Table 5 shows the output of the ANOVA analysis and whether we have a statistically significant difference between our group means on different number of groups with 5YD datasets. Therefore, all the amounts of group with six quality indicators have statistically significant difference in our patent clustering.

Fig. 8 shows the distribution of patent applications of 13 patent offices and 3 quality groups in 5 yearly datasets. We focus on top six major patent offices in which the patent publications of CN,

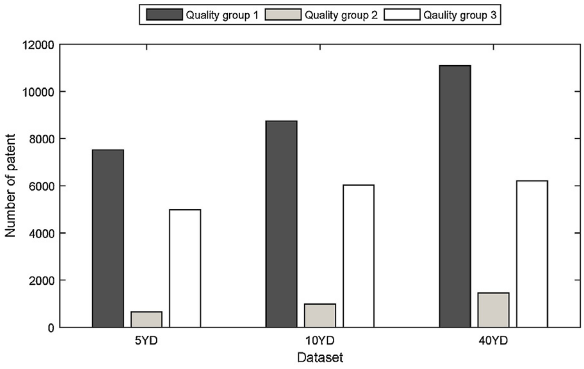

Fig. 5. Distribution of patent applications in three quality groups.

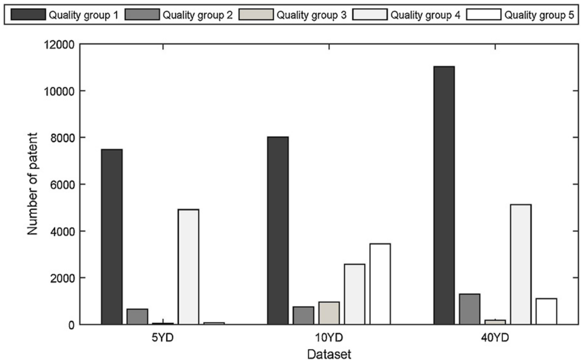

Fig. 6. Distribution of patent application in five quality groups.

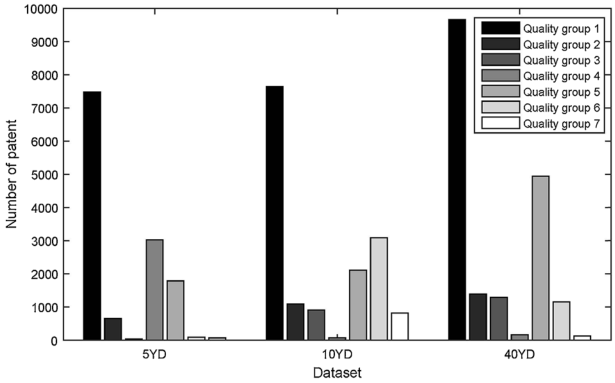

Fig. 7. Distribution of patent applications in seven quality groups.

Table 4
The quality summary on different groups with different time periods.

Amount of group	Quality group	5YD	10YD	40YD
3 QG	G1	0.0776	0.0720	0.0768
	G2	0.5419	0.5136	0.5050
	G3	1.1213	1.0822	1.0780
5 QG	G1	0.0744	0.0632	0.0749
	G2	0.5399	0.1896	0.4650
	G3	0.6569	0.5074	0.6872
	G4	1.1097	1.0009	1.0636
	G5	1.9142	1.1427	1.1421
7 QG	G1	0.0744	0.0595	0.0607
	G2	0.5396	0.1570	0.1801
	G3	0.6650	0.4862	0.4655
	G4	1.0404	0.7584	0.7060
	G5	1.2028	0.9740	1.0605
	G6	1.5689	1.1149	1.1077
	G7	1.9182	1.2391	1.4678

Table 5
ANOVA of different amounts of quality group in 5YD dataset.

Amount of group	Measure	Quality indicators					
		PQI1	PQI2	PQI3	PQI4	PQI5	PQI6
3 QG	Mean squares	6607.786	102255.224	21272266.570	66224.078	13522.313	1710.560
	F-value	174.717*	365.838*	1326242.283*	4715.469*	28.199*	26.842*
5 QG	Mean squares	3313.964	841364.292	10636132.180	34256.479	18978.022	4144.875
	F-value	87.619*	21387.040*	663006.425*	2500.79*	39.879*	66.068*
7 QG	Mean squares	42829.500	573144.778	7091550.496	33110.771	17002.250	3003.930
	F-value	2218.647*	16973.901*	452213.976*	3672.657*	35.871*	47.958*

* Between groups significantly different ($p < 0.05$).

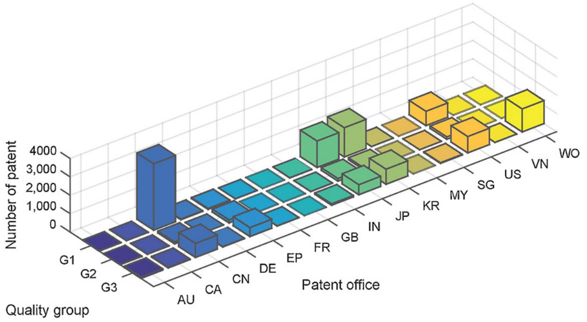

Fig. 8. Distribution of patent applications of 13 patent offices and 3QG in 5YD dataset.

JP and KR patent offices have the most patents in lower quality group (G1), but they have some of patents in high-quality group (G3). The EP and WO patent offices have the most patents in higher quality group (G3); The USA has the most patents in lower and higher quality groups.

In addition, we computed the number of IPC cross quality groups on IPC class 3 and IPC class 4 with 3QG and 5YD dataset. Table 6 shows that the average number of IPC class 3 on G1, G2 and G3 are 1.5343, 1.5690 and 1.7172, respectively. The average number of IPC class 4 on G1, G2 and G3 are 1.7387, 1.9878 and 2.048, respectively. Therefore, the higher quality groups of G3 in IPC class 3 and IPC class 4 have more number of IPC class and it means that the patent's techniques have very wide application in many industries.

4.3. Results on patent quality classification

The patent analysis by SOM model with quality indicators has defined the quality type for each patent of thin film solar cell. The patents for training and testing are using different amounts of quality groups (i.e. 3QG, 5QG and 7QG) with different period datasets (i.e. 5YD, 10YD, 40YD). All yearly datasets on testing data are of same period in 2013 and the training data period of 5YD dataset is from 2009 to 2012, 10YD dataset is from 2004 to 2012, and 40YD dataset from 1974 to 2012. The details of destruction of patent applications are shown in Table 7.

Table 8 shows the explanations on different feature percent on Gaussian kernel (Gauss.) and Polynomial kernel (Poly.). In Gaussian kernel, over 25% of components have 47.38%, 48.43% and 48.88% explained in all datasets; over 50% of components have 75.04%, 74.03% and 72.48% explained in all datasets; over 75% of components have 94.14%, 94.05% and 95.61% explained in all datasets. In Polynomial kernel, over 25% of components have 91.08%, 95.40% and 92.26% explained in all datasets; over 50% of components have 99.98%, 100% and 99.99% explained in all datasets; over 75% of components have 100% explained in all datasets. Finally, all 100% of components have 100% explained in all datasets, number of groups and kernels. The Polynomial kernel has higher variation explained when using less percent of feature.

Table 6
Statistics of number of IPC class on 3 quality groups and 5YD.

Quality group	**Average number of IPC class/per patent**	
	IPC class 3	IPC class 4
G1	1.5343	1.7387
G2	1.5690	1.9878
G3	1.7172	2.0748

Table 9 shows the classification performance by our proposed SOM-KPCA-SVM in the results of nonlinear feature extraction by Gaussian kernel. In the three groups with 5YD dataset, 75% KPCA feature has high accuracy and recall of 82.52% and 56.00% but not the best precision; with 10YD dataset, the 75% KPCA feature has high accuracy and recall of 82.83% and 65.34 but not the best precision; with 40YD dataset, the 100% KPCA feature has high accuracy is 82.35% but not the best precision and recall. In the five groups with 5YD dataset, the 75% KPCA feature has high accuracy and recall of 82.27% and 34.93% but not the best precision; with 10YD dataset, the 50% KPCA feature has high accuracy and precision of 72.06% and 47.49% but not the best recall. In the five groups with 40YD dataset, 75% KPCA feature has high accuracy of 82.83% but not the best precision and recall; with 5YD dataset, the 75% KPCA feature has high accuracy of 74.32% but not the best precision and recall; with 10YD datasets, the 50% KPCA feature has high accuracy of 71.38% but not the best precision and recall; with 40YD dataset, the 50% and 75% KPCA features have a high accuracy of 77.92%.

Table 10 shows that the results of nonlinear feature extraction by Polynomial kernel. In the three groups with 5YD dataset, the 50% KPCA feature has high accuracy and recall are 84.13% and 56.77% but not the best precision; with 10YD dataset, the 50% KPCA feature has high accuracy, precision and recall of 84.08%, 73.28% and 58.59%; with 40YD dataset, the 50% KPCA feature has high accuracy, precision and recall are 84.30%., 48.31 and 43.80%. In the five groups with 5YD dataset, the 50% KPCA feature has high accuracy and recall are 83.71% and 37.20% but not the best precision; with 10YD dataset, the 100% KPCA feature has high accuracy and recall are 72.00% and 48.45 but not the best precision; 40YD dataset, the 50% KPCA feature has high accuracy of 82.72% but precision and recall is not best. In the seven groups with 5YD dataset, the 50% KPCA feature has high accuracy and recall of 76.67% and 31.40% but not the best precision; with 10YD dataset, the 100% KPCA feature has high accuracy and recall of 71.24% and 41.84% but not the best precision; with 40YD dataset, the 50% KPCA features has high accuracy of 80.54% but not the best precision and recall. Table 11 shows our proposed SVM with KPCA by Polynomial kernel has the best performance compared with Gaussian kernel. Only one testing of the Polynomial kernel is not better than Gaussian kernel on five groups with 10YD dataset.

Table 12 shows that two classifiers' performances on accuracy measure. The SVM classifier in all cases is better than linear discriminant analysis (LDA), decision trees (DT) and artificial neural network (ANN). The overall average accuracies on four classifiers of SVM, LAD, DT and ANN are 79.79%, 72.66%, 71.41% and 60.79%, respectively.

Table 7
Patent application distribution of training and testing data.

Amount of group	No. of group	Training data			Testing data		
		5YD (2009–2012)	10YD (2004–2012)	40YD (1974–2012)	5YD (2013)	10YD (2013)	40YD (2013)
3 groups	G1	5,299	6,523	8,869	2,222	2,222	2,222
	G2	567	895	1,361	92	92	92
	G3	3,762	92	4,981	1,222	1,222	1,222
5 groups	G1	5,268	5,843	8,808	2,211	2,178	2,221
	G2	566	707	1,215	92	44	88
	G3	31	867	181	11	92	5
	G4	3,706	2,026	3,988	1,206	547	1,135
	G5	57	2,772	1,019	16	675	87
7 groups	G1	5,269	5,530	7,511	2,211	2,115	9,665
	G2	566	985	1,328	92	106	1,395
	G3	30	825	1,204	11	88	1,292
	G4	2,192	72	158	836	5	163
	G5	1,440	1,603	3,829	349	508	4,946
	G6	76	2,472	1,071	21	619	1,155
	G7	55	728	110	16	95	131

Table 8
The feature explained (%) of KPCA on different percent of feature.

Percent of KPCA feature	5YD		10YD		40YD	
	Gauss.	Poly.	Gauss.	Poly.	Gauss.	Poly.
25%	47.38	91.08	48.43	95.40	45.88	92.26
50%	75.04	99.98	74.03	100.00	72.48	99.99
75%	94.14	100.00	94.05	100.00	94.61	100.00
100%	100.00	100.00	100.00	100.00	100.00	100.00

Table 9
The performance (%) on different nonlinear features by Gaussian kernel.

Amount of group	KPCA feature	5YD			10YD			40YD		
		ACC	PRS	REC	ACC	PRS	REC	ACC	PRS	REC
3 QG	25%	65.78	57.84	39.50	65.36	61.00	38.74	64.79	45.29	35.62
	50%	79.75	73.72	52.19	81.67	64.07	54.44	81.59	60.76	53.26
	75%	**82.52**	66.68	56.00	**82.83**	59.08	56.34	82.32	59.56	54.60
	100%	82.10	60.29	55.28	82.21	58.15	55.67	**82.35**	55.24	54.05
5 QG	25%	65.72	57.78	23.83	61.45	39.63	30.43	62.81	62.81	20.00
	50%	79.24	72.93	31.33	**72.06**	47.49	46.77	79.86	61.76	41.61
	75%	**82.27**	40.88	34.93	71.63	42.36	46.79	**80.20**	48.86	42.37
	100%	81.70	38.55	34.42	72.00	42.14	47.45	80.12	46.05	46.39
7 QG	25%	62.53	62.53	14.29	61.57	34.89	26.57	61.88	35.57	26.32
	50%	73.59	46.61	26.64	**71.38**	43.44	39.14	**77.97**	56.03	37.96
	75%	**74.32**	36.20	28.91	70.87	41.99	41.22	**77.97**	42.49	38.01
	100%	73.42	36.86	29.41	71.24	44.54	41.84	77.80	51.06	41.24

Table 10
The performance (%) on different nonlinear features by Polynomial kernel.

Amount of group	KPCA feature	5YD			10YD			40YD		
		ACC	PRS	REC	ACC	PRS	REC	ACC	PRS	REC
3 QG	25%	82.58	80.89	55.59	83.94	62.44	57.05	67.65	32.98	30.03
	50%	**84.13**	60.76	56.77	**84.08**	73.28	58.59	**84.30**	48.31	43.80
	75%	82.44	73.51	55.90	82.92	59.4	56.39	82.15	44.37	41.03
	100%	82.07	60.27	55.26	82.21	58.15	55.67	82.35	41.68	40.78
5 QG	25%	81.48	64.21	33.62	70.76	42.85	42.23	62.81	62.81	20.00
	50%	**83.74**	58.80	37.20	70.79	44.73	46.19	**82.72**	62.12	45.56
	75%	82.15	42.36	34.84	71.38	42.09	46.66	80.18	49.86	42.55
	100%	81.70	38.55	34.42	**72.00**	42.14	48.45	80.12	46.05	46.39
7 QG	25%	75.45	56.14	29.05	69.85	44.55	37.70	64.17	46.96	29.57
	50%	**76.67**	53.80	31.40	70.62	39.21	39.96	**80.54**	46.76	40.59
	75%	73.78	36.70	28.58	71.10	43.19	41.52	77.97	42.76	38.10
	100%	73.44	36.87	27.42	**71.24**	44.54	41.84	77.80	51.06	41.24

Table 11
The accuracy comparison (%) on Gaussian kernel (Gauss.), Polynomial kernel (Poly.) and unused KPCA (Non).

Amount of group	5YD			10YD			40YD		
	Non	Gauss.	Poly.	Non	Gauss.	Poly.	Non	Gauss.	Poly.
3 QG	82.10	82.53	**82.58**	82.21	82.83	**84.08**	82.35	82.32	**84.30**
5 QG	81.73	82.27	**83.74**	**72.06**	**72.06**	72.00	80.18	80.20	**82.72**
7 QG	73.19	74.32	**76.67**	**71.44**	71.38	71.24	77.80	77.97	**80.54**

Table 12
The accuracy comparison (%) on SVM, LDA, DT and ANN classifiers.

Amount of group	5YD				10YD				40YD			
	SVM	LDA	DT	ANN	SVM	LDA	DT	ANN	SVM	LDA	DT	ANN
3 QG	**82.58**	79.53	78.47	73.33	**84.08**	78.25	75.71	73.28	**84.30**	74.55	77.49	69.37
5 QG	**83.74**	79.10	77.26	64.34	**72.06**	66.32	64.37	50.79	**82.72**	70.96	72.94	62.87
7 QG	**76.67**	69.85	68.07	58.82	**71.44**	66.37	61.20	43.18	**80.54**	69.01	67.17	51.16

4.4. Discussions and managerial implications

Finding out that patent analysis and classification helped companies and industry to understand current patent quality and future patent quality. There are four directions for the discussion:

(1) First, the differences in the number of groups will impact analysis rationality. The groups have too high similarity among quality groups when they used three clustering, thus differences of groups are very low. We suggest the number of groups to be not more than seven because the number of patents in seven groups clustering is too less among groups.
(2) Second, the nonlinear feature transformation has improved classification performance. In this case, the Polynomial kernel has a lot of accuracy, precision and recall compared with Gaussian kernel. But the process time of Gaussian kernel is very less.
(3) Third direction, according to the result of Fig. 8 that there are three major patent office belonging to the higher quality group which is EP, US and WO. The CN has 80.80% patents belonging to the lower quality group because it has 33% patents of all offices in the last five year patent data. The patents of WO patent office in 5YD dataset are all belonged to higher quality because most of high valuable or potential patents are applied in WO patent by inventors.
(4) Fourth direction, our proposed quality classification system determined the quality type for new application or publication patent. Thus, the system can help R&D to think about the benefit of high-quality patent on new production. In addition, the automatic quality analysis and classification can avoid overreliance manpower, excessive process time and too slow reaction to market trends.

5. Conclusions

This study proposed three data mining approaches to patent analysis and patent quality forecasting. The SOM-KPCA-SVM patent quality system combined self-organizing maps, kernel principal component analysis and support vector machine to classify patent quality of a thin film solar cell in solar industry. The SOM has successful cluster patent into different quality groups and its result has statistically significant difference in the quality indicators between the quality groups. The KPCA has effectively transformed a nonlinear feature space from characteristics of a patent document and it can improve classification performance. The SVM has built a powerful classification model for patent quality problem. Therefore, our proposed SOM-KPCA-SVM automatic patent quality classification system has improved analysis time, cost and manpower by traditional patent analysis approaches. Thus, SOM-KPCA-SVM system can take a short time to determine patent quality. Through our automatic quality classifier, we can promote precision of decision-making for R&D of companies or industry. In addition, our analysis and classification used the larger patent data that avoid mark analysis mistake from less data for decision-making. In the future research work, we will consider the degree of patent quality. Each patent will assume a degree of quality instead of same quality type. The detailed quality level can provide more precise quality information for decision-making.

References

[1] A.J.C. Trappey, C.V. Trappey, C.Y. Wu, C.L. Lin, A patent quality analysis for innovative technology and product development, Adv. Eng. Informatics 26 (1) (2012) 26–34.
[2] A.J.C. Trappey, C.V. Trappey, C.Y. Wu, C.Y. Fan, Y.L. Lin, Intelligent patent recommendation system for Innovative design collaboration, J. Network Comp. Appl. 36 (6) (2013) 1441–1450.
[3] Abbas, L. Zhang, S.U. Khan, A literature review on the state-of-the-art in patent analysis, World Patent Info. 37 (2014) 3–13.
[4] T. Dereli, A. Durmuşoğlu, Classifying technology patents to identify trends: Applying a fuzzy-based clustering approach in the Turkish textile industry, Technol. Soc. 31 (3) (2009) 263–272.
[5] F. Narin, Patent bibliometrics, Scientometrics 30 (1) (1994) 147–155.
[6] F. Narin, K. Hamilton, D. Olivastro, The increasing linkage between US technology and public science, Res. Policy 26 (1997) 317–330.
[7] J.R. Allison, M.A. Lemley, K.A. Moore, R.D. Trunkey, Valuable patents, Georgetown Law J. 92 (2004).
[8] J. Segev, Kantola, Identification of trends from patents using self-organizing maps, Expert Syst. Appl. 39 (18) (2012) 13235–13242.
[9] P.C. Chang, J.L. Wu, A critical feature extraction by kernel PCA in stock trading model, Soft Comput. 19 (5) (2015) 1393–1408.
[10] P.C. Chang, J.L. Wu, The weighted support vector machines for the stock turning point prediction, in: in: Proceedings of the Intelligent System design and Applications (ISDA), Okinawa, Japan, 2014, pp. 205–210.
[11] S. Ercan, G. Kayakutlu, Patent value analysis using support vector machines, Soft Comput. 18 (2) (2014) 313–328.
[12] W.D. Yu, S.S. Lo, Patent analysis-based fuzzy inference system for technological strategy planning, Automat. Constr. 18 (2009) 770–776.
[13] S.S. Ju, M.F. Lai, C.Y. Fan, Using patent analysis to analyze the technological developments of virtualization, Procedia - Social Behav. Sci. 57 (2012) 146–154.
[14] S. Brügmann, N. Bouayad-Agha, A. Burga, S. Carrascosa, Towards content-oriented patent document processing: Intelligent patent analysis and summarization, World Patent Info. 40 (2015) 30–42.
[15] M. Grimaldi, L. Cricelli, M.D. Giovanni, F. Rogo, The patent portfolio value analysis: A new framework to leverage patent information for strategic technology planning, Technol. Forecast. Soc. Change 94 (2015) 286–302.
[16] J. Bessen, The value of U.S. patents by owner and patent characteristics, Res. Policy 37 (2008) 932–945.
[17] S.S. Nair, M. Mathew, D. Nag, Dynamics between patent latent variables and patent price, Technovation 31 (2011) 648–654.
[18] M.D. Saint-Georges, B.P. Potterie, A quality index for patent systems, Res. Policy 42 (2013) 704–719.
[19] E. Archontopoulos, Prior art search tools on the Internet and legal status of the results: a European Patent Office perspective, World Patent Info. 26 (2) (2004) 113–121.
[20] E.S. Simmons, B.D. Spahl, Of submarines and interference: legal status changes following citation of an earlier US patent or patent application under 35 USC §102 (e), World Patent Info. 22 (3) (2000) 191–203.
[21] M. Hirschey, V.J. Richardson, Valuation effects of patent quality: A comparison for Japanese and U.S. firms, Pacific-Basin Fin. J. 9 (2001) 65–82.
[22] J. Dang, K. Motohashi, Patent statistics: A good indicator for innovation in China? Patent subsidy program impacts on patent quality, China Econ. Rev. 35 (2015) 137–155.
[23] Y.S. Chen, K.C. Chang, The relationship between a firm's patent quality and its market value-The case of US pharmaceutical industry, Technol. Forecast. Soc. Change 77 (2010) 20–33.
[24] R. Frietsch, P. Peter Neuhäusler, T. Jung, B.V. Looy, Patent indicators for macroeconomic growth-the value of patents estimated by export volume, Techonovation 34 (2014) 546–558.
[25] D. Harhoff, F.M. Scherer, K. Vopel, Citations, family size, opposition and the value of patent rights, Res. Policy 32 (2003) 1343–1363.
[26] D. Deng, N. Kasabov, On-line pattern analysis by evolving self-organizing maps, Neurocomputing 51 (2003) 87–103.
[27] P. Juntunen, M. Liukkonen, M. Lehtola, Y. Hiltunen, Cluster analysis by self-organizing maps: An application to the modelling of water quality in a treatment process, Appl. Soft Comput. 13 (7) (2013) 3191–3196.
[28] S. Bouhouche, M. Yahi, J. Bast, Combined use of principal component analysis and self organisation map for condition monitoring in pickling process, Appl. Soft Comput. 11 (3) (2011) 3075–3082.

[29] F.A. Ravalison, N. Rabenja, Using patent statistics and principal component analysis to predict global competition, Int. J. Ind. Eng. Manage. 2 (2) (2011) 34–50.

[30] R. Shao, W. Hu, Y. Wang, X. Qi, The fault feature extraction and classification of gear using principal component analysis and kernel principal component analysis based on the wavelet packet transform, Measurement 52 (2014) 118–132.

[31] V. Vapnik, The nature of statistical learning theory, 2nd ed., Springer, New York, 1999.

[32] V. Cortes, Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297.

[33] C.H. Wu, Y. Ken, T. Huang, Patent classification system using a new hybrid genetic algorithm support vector machine, Appl. Soft Comput. 10 (4) (2010) 1164–1177.

[34] C.Y. Chiu, P.T. Huang, Application of the honeybee mating optimization algorithm to patent document classification in combination with the support vector machine, Int. J. Automat. Smart Technol. 3 (3) (2013) 179–191.

[35] S. Venugopalan, V. Rai, Topic based classification and pattern identification in patent, Technol. Forecast. Soc. Change 94 (2015) 236–250.