A Patent Quality Classification System Using a Kernel-PCA with SVM

Pei-Chann Chang
Innovation Center for Big Data & Digital Convergence and Dept. of Information Management
Yuan Ze University
Taoyuan Taiwan
iepchang@saturn.yzu.edu.tw

Jheng-Long Wu
Innovation Center for Big Data & Digital Convergence and Dept. of Information Management
Yuan Ze University
Taoyuan Taiwan
vickeytsao@gmail.com

Cheng-Chin Tsao
Innovation Center for Big Data & Digital Convergence and Dept. of Information Management
Yuan Ze University
Taoyuan Taiwan
vickeytsao@gmail.com

Meng-Hsuan Lin
Innovation Center for Big Data & Digital Convergence and Dept. of Information Management
Yuan Ze University
Taoyuan Taiwan
syuan417@gmail.com

***Abstract*—Data mining (DM) approaches such as clustering and classification are employed in this paper to identify and classify the patent quality. We develop an effective and automatic patent quality classification system. First, the Self-organizing map (SOM) is used to cluster patents automatically into different quality groups with patent quality indicators instead of via expert identification. Then, the Kernel principal component analysis (kernel-PCA) is used to extract key indicator to improve classification performance. Finally, the Support vector machine (SVM) is used to build the quality classification model. The proposed classification model is applied to classify patent quality automatically in solar industries. Experimental results show that our proposed approach KPCA-SVM can improve the performance of the patent quality classification when compared with the traditional method. Another advantage is that the computational time is largely reduced.**

Keywords- patent quality classification, self-organizing maps, support vector machine, kernel principal component analysis, solar industries.

I. INTRODUCTION

An important issue of patent analysis is patent quality. 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]. In addition, traditional patent analysis requires spending much time, cost and manpower. Therefore, the potential for high quality patent determining approach need to shorten the time for business or production. Recently, the self-organizing map (SOM) is used to analysis patent for patent trend [4] and regional innovation systems [5]. It can analyse patent current situation and trend, but it doesn't provide a solution for determine future patent quality. The future patent recognition is a key research for present time because patent impact on the industry. The future patent recognition is a key research for the time because patent impact on the industry need to response quickly. Therefore, support vector machine (SVM) forecasting model can solve patent classification problem for predict future unknown patent classification such as quality [6]. In this study, we propose a KPCA-SVM patent quality classification system that combines three data mining (DM) methods such as SOM, Kernel-PCA and SVM. The SOM is used to cluster patents into several groups for qualities according to their data characteristics. Then, the quality indictor can compute the quality levels for delimit patents quality on each groups. The kernel principle components analysis (kernel-PCA) is based on principle components analysis and used to transform patent data into new features set by nonlinear kernel mapping. SVM is used to build classification model for patent quality problem. This methodology helps experts rank and set values on patent quality in solar industry. Therefore, a trained model can evaluate unknown patents' quality, better enable engineers and product designers forecast patent potential for product development.

II. LITERATURE REVIEWS

A. Patnet Analysis

There are various tools utilized by organizations for analyzing patents. These tools are capable of performing wide range of tasks, such as forecasting future technological trends, detecting patent infringement and determining patent quality and so on [3]. 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]. The eventually may result in imprecise recommendation of patents. However, the larger data and indicators for patent quality forecasting are needed.

B. Patnet Quality Indicator

The primary patent quality indicators are related to investment, maintenance, and litigation, which form a basis for assessing patent quality, when the evaluation focuses on the potential patents for business. One kind of indicator of patent quality is legal status (LS) that it can show which technologies are hot and which are not for business intelligence. The legal status and search tools on the internet are very sensitive that emphasis given to the issues related to the date of availability to the public of an Internet disclosure, its conformance and its possibly non-prejudicial nature [7]. The legal status change that suggestions on how these changes may be tracked are provided, specifically resolution

 ISBN: 978-1-61208-419-0

is also complicated by the "first to invent" concept in US patent law [8].

C. 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 group according to similarities and patterns found in the data set, using some form of distance measure which use 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. The authors [4] used SOM to identify patent trends that they analyze patent knowledge to identify research trends. They tested on patents from the United States Patent and Trademark Office (USPTO) and result both an overview of the directions of the trends and a drill-down perspective of current trends. Another algorithm using SOM that is evolving self-organizing map (ESOM), which features an evolving network structure and fast on-line learning. Their result shows that ESOM achieved better or comparable performance with a much shorter learning process [9].

D. Kernel Principle Component Analysis

Principal component analysis (PCA) is very useful to extract nonlinear features for many research applications. The Kernel-PCA is an extension of PCA using the kernel mapping before the Eigen-problem. Kernel-PCA is as a nonlinear alternative to classical PCA of combustion composition space is investigated. PCA is mathematically defined. The PCA is widely used in many field researches. The research [1] wants to identify the key impact factors using PCA and they selected a lot of variables according to first five components indicate. The Kernel-PCA is used to critical feature extraction in stock trading model and capture best performance compared to PCA, ICA and so on [10].

E. Support Vector Machinie

Support vector machine is a machine learning algorithm and widely used for classification problems [5]. 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. Support vector machine map input vectors into the high dimensional feature space via the non-linear mapping. An effective decision hyper-plane is developed to distinguish the correct training data. An approach is proposed integrated with a hybrid genetic-based support vector machine (HGA-SVM) model for developing a patent classification system [11]. But they are needed expert's knowledge to analysis. The authors claim that they use these models in real-world cases of patent classification rather than only use for International Patent Classification (IPC). The study integrated the honey-bee mating optimization algorithm with SVM (HBMOSVM) for patent document categorization. In their results show that the HBMOSVM could result in better patent documentation accuracy and better F-measure performance as an evaluation index than GASVM model in patents document categorization [12].

III. PROPOSED METHODOLOGY: KPCA-SVM PATENT QUALITY CLASSIFICATION SYSTEM

This study proposes an automatic patent quality classification system that integrated methodology as KPCA-SVM; the components used are SOM, kernel-PCA and SVM approaches. Fig. 1 shows that, first, we collect the patent data related industries from the patent database, use SOM approach to cluster patens into several groups and use patent quality indicators to compute potential quality on each group for quality identification; second, the kernel-PCA extracts key indicators into nonlinear feature space; finally, SVM forecasting model is used to build patent quality classification model using nonlinear feature space by kernel-PCA in order to predict quality of future patent who has potential effectiveness. Then, we evaluate patent quality classification system and forecast quality level for each paten by our proposed system. Our system is developed as follows:

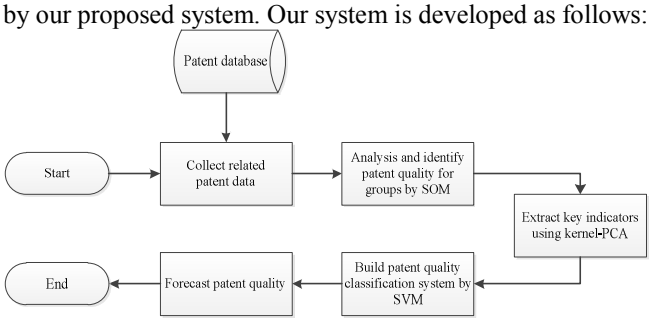

Figure 1. The system framework of KPCA-SVM.

A. Patent Quality Analysis and Identification by SOM with Quality Indicators

The SOM approach is adopted to cluster the patents, to identify patent quality and to explore hidden patterns among these patents. According to SOM, these patents will be split into several groups (clusters) and each cluster includes number of patents with a similar patent quality. This SOM 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. Details of the SOM algorithm are listed as follows:

- *Step 1:* Set-up the parameters in the SOM network.
- *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.
- *Step3:*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* such as

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

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

- *Step 4:* Adjust the weight of winning neuron *c* and all neighbor units.

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

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 $ch_{ci}(t)$ defines the region of influence that the input pattern has on the SOM and consists of two parts: the neighborhood function $h(\|\cdot\|, t)$ and the learning rate function α^t,

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

where *r* is the location of the neuron on two dimensional map grids. 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

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

- **Step 5:** repeat step 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.

The patents of each group will compute the quality by quality indicators such as legal status. The quality levels of each group are calculated as follows:

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

where q_{ij} denotes value of quality of *j*th variable of *i*th patent in *g*th group.

B. Extracting Key Patent Indicators by Kernel-PCA

In order to compute dot products of the form, we use kernel representation of the form.

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

which allows us to compute the value of the dot product in *F* without having to carry out the map Φ. A number of kernel functions exist as been chosen before we apply the algorithm. The representative Gaussian kernel function *K* is described as follows:

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

Given a set of *m*-dimensional normalized patent indices $x_k \in R^m$, we compute the kernel matrix $K \in R^{N \times N}$ from two kernel methods,

$$K_{ij} = (\Phi(x_i), \Phi(x_j)) = [k(x_i, x_j)] \tag{9}$$

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

$$K' = K - C*K - K*C + C*K*C \tag{10}$$

C. Building Patent QualityClassification Model by SVM

The patent data of identified quality will be split into two datasets, i.e., training data set and testing data set. The new feature spaces of training data *training*(tr^k) and testing data *testing*(ts^k) are represented by the $\tilde{\Phi}(x)$ and α_i^k. For patent variables *x* in training period, we extract a nonlinear component via

$$\begin{aligned} Training\ (tr^k) &= (v^k, \tilde{\Phi}(x)) = \sum_{i=1}^{N} \alpha_i^k (\tilde{\Phi}(x_i), \tilde{\Phi}(x)) \\ &= \sum_{i=1}^{N} \alpha_i^k \tilde{K}(x_i, x) \end{aligned} \tag{11}$$

where $\tilde{\Phi}(x)$ is the mean centered.

The testing data is unknown data as well as the future data. We cannot directly use the testing data to compute the mean centering $\Phi(x)$ and eigenvalues α in PCA processes. In order to avoid this problem, we use the $\tilde{\Phi}(x)$ and α_i^k from the training data to extract a nonlinear component.

$$\begin{aligned} Testing\ (ts^k) &= (v^k, \tilde{\Phi}(x)) = \sum_{i=1}^{N} \alpha_i^k (\tilde{\Phi}(y_i), \tilde{\Phi}(x)) \\ &= \sum_{i=1}^{N} \alpha_i^k \tilde{K}(y_i, x) \end{aligned} \tag{12}$$

 ISBN: 978-1-61208-419-0

where y denotes the normalized variables $y_k \in R^m$ in testing data.

The input vector of SVM training employ the new nonlinear feature space *training*(tr^k) by kernel-PCA and the output vector of quality level is given by SOM with quality indicators. The trained model of SVM is then used to evaluate the testing data *testing*(tr^k) for patent quality forecasting. Thus, the proposed model can be applied to build the patent quality model for evaluating the potential of patents.

IV. EXPERIMENTAL RESULTS

In this paper, these patents were divided into seven groups in order to discriminating quality levels into seven degrees. The parameters of SOM including epochs and cluster are set up to 5,000 and 7, respectively. In SVM model, the kernel is radial basis function (RBF), while the cost is 256 and gamma is 0.25. The patent data of solar industry are collected from patent database of Thomson Innovation which has 60,000 patents in eleven patent offices. Seven different quality groups are used in SOM and its quality levels are sorted by legal status of quality indictor. The Table 1 shows that the highest legal status is in group 7, the second is in group 6 and the lowest is in group 1. The legal status on group 7 is 10.120 and the number of patents is 2,235. We observed that the higher of patent quality is, the less of patent number is. The other way around is for low patent quality patent.

TABLE I. THE LEGAL STATUS STATISTICS ON EACH GROUP

	Group						
	G1	*G2*	*G3*	*G4*	*G5*	*G6*	*G7*
No. of patent	22,431	6,249	9,796	4,564	11636	3,089	2,235
Avg. of legal status	1.90	2.15	2.35	3.41	3.53	9.06	10.10

Fig. 2 shows that the distribution on patent applications for 11 patent offices. The two larger shares of patents are US and CN. Their patents exist in different quality levels because their markets are the most important economic region attracting many patent applications.

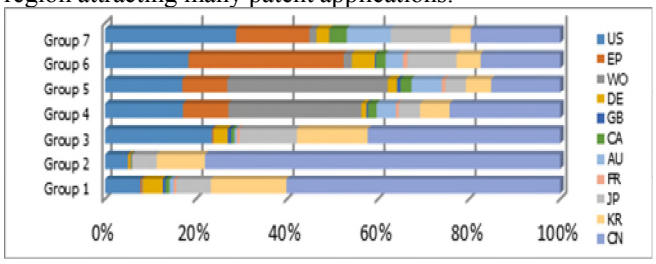

Figure 2. Distribution on patent applications for 11 patent offices

We focus on group 7 since the patent offices of US; CN; EP; JP and AU include 88% shares as shown in Fig. 3. Other offices are much less involved in solar industry. In addition, the patent applies in EP patent offices are high quality which are in Group 4, 5, 6 and 7. However, the quality levels are evenly spread in CN patent offices.

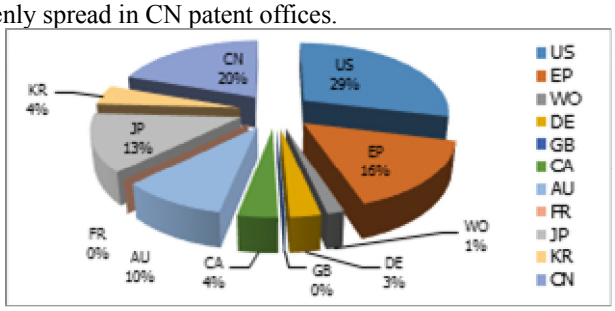

Figure 3. Distribution on patent applications on highest quality group (Group 7)

The results show the forecasting performance for KPCA-SVM and decision tree (DT). The optimal parameters were decided from the training data for kernel-PCA and SVM. Table 2 shows that our proposed model has best performance on three measure indicators and they are 99.71% on accuracy, 99.28% on average precision of all classes, and 99.31% on average recall of all classes.

TABLE II. FORECASTING PERFORMANCE FOR DIFFERENT CLASSIFICATION MODEL

Classifier	Accuracy	Precision	Recall
KPCA-SVM	99.91%	99.28%	99.31%
DT	96.86%	95.42%	95.08%

V. CONCLUSIONS

We proposed the KPCA-SVM patent quality system which combined SOM, kernel-PCA and SVM data mining approaches in solar industry. The experimental results showed that the proposed approach has a better performance when compared with other traditional approaches in terms of time consuming, cost and manpower. The proposed approach take a shorten time to determine patent quality and has 99.91% accuracy. In addition, the proposed system performs even better for a larger patent data and making fast and accurate recommendations. In the future research work, we will consider the relationship between patent quality and patent value creation. Therefore, different patent quality can be closely related to different patent values via accurate classification system.

REFERENCES

[1] Trappey, A.J.C., Trappey, C.V., Wu, C.Y. & Lin, C.L, A patent quality analysis for innovative technology and product development, Adv Eng Inform, 26(1), pp. 26-34, 2012.

[2] Trappey, A.J.C., Trappey, C.V., Wu, C.Y.W., Fan, C.Y. & Lin, Y.L., Intelligent patent recommendation system for Innovative design collaboration, J Netw Comput Appl, 36, pp. 1441-1450, 2013.

[3] Abbas, A., Zhang, L. & Khan, S.U., A literature review on the state-of-the-art in patent analysis, World Pat Inf, 37, pp. 3-13, 2014.

 ISBN: 978-1-61208-419-0

[4] Segev, A. & Kantola, J., Identification of trends from patents using self-organizing maps, Expert Syst Appl, 39(18), pp. 13235-13242, 2012.

[5] Hajek, P., Henriques, R. & Hajkova, V., Visualising components of regional innovation systems using self-organizing maps—Evidence from European regions, Technol Forecast Soc Change, 84,pp.197-214, 2014.

[6] Ercan, S. & Kayakutlu, G., Patent value analysis using support vector machines, Soft Comput, 18, pp. 313-328, 2014.

[7] Archontopoulos, E., Prior art search tools on the Internet and legal status of the results: a European Patent Office perspective, World Pat Inf, 26(2), pp. 113-121, 2004.

[8] Simmons, E.S. & Spahl, B.D., Of submarines and interference: legal status changes following citation of an earlier US patent or patent application under 35 USC §102 (e), World Pat Inf, 22(3), pp. 191-203, 2000.

[9] Deng, D. & Kasabov, N., On-line pattern analysis by evolving self-organizing maps, Neurocomputing, 51, pp. 87-103, 2003.

[10] Chang, P.C. & Wu, J.L., A critical feature extraction by kernel PCA in stock trading model, Soft Comput., DOI 10.1007/s00500-014-1350-5.

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

[12] Chiu, C.Y., & Huang, P.T. Application of the honeybee mating optimization algorithm to patent document classification in combination with the support vector machine, Int. j. autom. smart technol, 3(3), pp.179-191, 2013.0.

 ISBN: 978-1-61208-419-0