Scientometrics (2010) 82:217–241
DOI 10.1007/s11192-009-0008-z

Discovery of factors influencing patent value based on machine learning in patents in the field of nanotechnology

Scott D. Bass · Lukasz A. Kurgan

Received: 18 September 2007 / Published online: 9 June 2009

Abstract Patents represent the technological or inventive activity and output across different fields, regions, and time. The analysis of information from patents could be used to help focus efforts in research and the economy; however, the roles of the factors that can be extracted from patent records are still not entirely understood. To better understand the impact of these factors on patent value, machine learning techniques such as feature selection and classification are used to analyze patents in a sample industry, nanotechnology. Each nanotechnology patent was represented by a comprehensive set of numerical features that describe inventors, assignees, patent classification, and outgoing references. After careful design that included selection of the most relevant features, selection and optimization of the accuracy of classification models that aimed at finding most valuable (top-performing) patents, we used the generated models to analyze which factors allow to differentiate between the top-performing and the remaining nanotechnology patents. A few interesting findings surface as important such as the past performance of inventors and assignees, and the count of referenced patents.

Keywords Patent · Patent value · Nanotechnology · Machine learning · Classification · Feature selection

Introduction

Machine learning refers to the ability of a machine to recognize patterns in data and to improve its performance based on experience that can be learned from these patterns. It encompasses a number of computational techniques that generate models of data, which can be used to find interesting relationships and that can be applied to provide useful insights, such as ability to predict certain facts, that concern the existing and the future data. As machine learning techniques are independent of the application domain, they

S. D. Bass · L. A. Kurgan (✉)
Department of Electrical and Computer Engineering, University of Alberta,
9701 116 Street, Edmonton, AB T6G 2V4, Canada
e-mail: lkurgan@ece.ualberta.ca

provide a valuable platform to analyze and learn from data in various disciplines and industries (Van Someren and Urbancic 2005). One of the main advantages of this approach is that it can generate compact, human-readable models from large amounts of data. In this project, we apply machine learning to generate and analyze models extracted from a large collection of patent records in a sample nanotechnology industry. Our main goal is to use these models to learn which factors, which are extracted from a patent record, allow to determine its future value.

Innovation plays a key role in economic development and is therefore a primary concern for practitioners, policy makers, and researchers (van Looy et al. 2006). Among other things, this has led to the theoretical and empirical analysis of patent value, which has attracted the attention of economists and government bodies for years (Rozhkov and Ivantcheva 1998; Reitzig 2003). Some studies have suggested the roles of certain actors and influences on innovation (e.g., Reitzig 2003); however, many of these roles are still not completely understood. One method that has been used in the past to study these roles and the paths of technological development and performance is the analysis of data found in patents (Verbeek and Debackere 2006), especially the data appearing on the cover sheet of each patent.

A patent is a document, containing structured, rich content regarding technological innovations that is accessible to the general public (Huang et al. 2003). Each patent document is issued by an authorized government agency and grants the owner a monopoly over the exploitation of a precisely defined technological advancement or incremental improvement (i.e., new device, apparatus, or process) over a stated period of time, e.g., 20 years in the United States (Connolly and Hirschey 1988; Debackere et al. 2002; Griliches 1990; Gupta 1999).

Patents permit the study of technological change since they represent inventive activity and output from applied research over different fields, countries, and time (Trajtenberg 1990; Hullmann and Meyer 2003). Since they are (1) an unchangeable written reflection of research and development output, (2) reveal trends in technology, and (3) provide considerable information about the applicants' approaches to the research, development, and marketing activities, their analysis helps not only study technological trends but also analyze science and technology as well as economic policies (Rozhkov and Ivantcheva 1998). This makes patent analysis (e.g., using statistics) an important tool for assessing the performance of technological systems (Wallin 2005). For example, patent counting, clustering, and citation analysis have been used to evaluate inventive activity at the corporate, industry, or national levels (Wallin, 2005). We also note that recent studies show increased activities related to commercialization of academic knowledge through patents, and thus academic patenting should be considered as another interesting dimension to perform patent analysis (Baldini and Grimaldi 2007; Leydesdorff and Meyer 2007).

The analysis of the information contained in patents is one of the most established and historically reliable methods of quantifying technological output (Debackere et al. 2002; Verbeek and Debackere 2006). It originated with the thought that the detailed information contained in patents might have a bearing on the importance of the innovations contained therein and that this information could be used to generate patent indicators that could act as proxies for the value of these innovations (Trajtenberg 1990). This has led to the related fields of patinformatics (Trippe 2003) or patent bibliometrics (Narin 1994). The term patinformatics encompasses the macro-level analysis of patent information for purposes such as patent intelligence, patent mapping, and patent citation analysis (Trippe 2003). Similarly, patent bibliometrics studies the mathematical and statistical patterns of citations in both scientific literature and patents, potentially to determine a patent's value (Wang

2007; Verbeek et al. 2002). For example, the count of patents can provide a basic indicator of the scientific and technological productivity whereas the count of citations to these patents can indicate the impact of the research as well as the linkage between science and technology (Narin and Hamilton 1996).

These fields have far reaching applications such as corporate assessments on the tactical and strategic levels, assessments of research and development programs and policies, intellectual property management, company valuation, competitive intelligence, modelling technological knowledge flows, economic modelling, technological trend evaluation, and modelling technological clusters (Breitzman and Mogee 2002; Gay and Le Bas 2005; Reitzig 2003; Lo 2008). In the area of intellectual property management, patent analysis can help focus efforts on patents that add the most value to a company since a company's patent portfolio is often skewed with a small proportion of the patents in the portfolio with high value and impact (Breitzman and Mogee 2002). In the area of competitive intelligence, patent analysis can be used in formulating industrial competitive strategies by evaluating the importance, technological strength, and creativity of both the company's patent portfolio and that of its competition (Albert et al. 1991; Chen et al. 2007a, b). Furthermore, it can be used to identify and characterize key companies in a technological area (Gupta 1999) and indicate the innovative and technological performance of high-tech companies (Hagedoorn and Cloodt 2003; Tong and Frame 1992). In the area of research and development programs and policies, patent analysis can indicate the relative success of a program or policy aggregated by technology, department, institution, region, or country (Connolly and Hirschey 1988; Narin and Hamilton 1996; Narin 1993). While in the area of valuation, patent analysis has found correlations between corporate performance (e.g., stock market valuations) and strong technology. For example, two models have been proposed for selecting stock portfolios based on patent indicators (Narin et al. 2004). Finally, in the area of trend evaluations, patent analysis has been used to identify key developments in the history of specific technological areas (Chen et al. 2007a, b).

Some of the advantages of using patent analysis and patent indicators as measures of technological activity include: (1) the proximity of patents to the inventive and innovative activities; (2) the range of fields covered by patents; (3) the geographical scope of patents; and (4) the accessibility and availability of patent data (Debackere et al. 2002). However, patent analysis is not without its shortcomings. For example, there are differences among the various patent systems (e.g., United States Patent and Trademark Office, European Patent Office, etc.) due to variations in legal, geographic, economic, and cultural factors. Also, there are variations in propensities to patent between different firms, technological fields, countries, etc. (Debackere et al. 2002). Despite these shortcomings, nothing compares to patent analysis in terms of the quality of data, accessibility, and detail (Debackere et al. 2002; Griliches 1990).

With patent analysis comes the question of how to assign a value to patents. While there is some debate about the definition of patent value, many patent analysis studies suggest using the backward citations as an index of the importance or value of patents (Gay and Le Bas 2005; Trajtenberg 1990). These backward citations are based on the front-page examiner citations (Trajtenberg 1990) from future patents that reference the current patent and are the basis for most patent citation analyses (Karki 1997). These are used since they are seen as pertinent to the subject matter of the patent (Wang 2007) and represent the 'shoulders' for which an invention is based (Gay and Le Bas 2005). However, there is some debate as to the meaning of the backwards citations and whether they imply technological value or economic value.

One hypothesis is that the backwards citations represent the technological value, quality or importance. In other words, a patent that is highly cited by subsequent inventions has a large technological value. Studies in this area have shown correlations between key or essential patents (i.e., those that generate the most value or impact) and the number of backward citations (e.g., Albert et al. 1991; Carpenter et al. 1981; Chen et al. 2007a, b; Debackere et al. 2002; Wang 2007). These studies treat citation counts as a proxy for the impact or technical relevance of the information or innovation found in the patent. The second hypothesis is that highly cited patents represent innovations that have high economic value since subsequent patents are the result of costly research and development efforts (Gay and Le Bas 2005). Studies in this area have shown correlations between the citation-weighted patent portfolios and market value of companies (Chen et al. 2007a, b; Debackere et al. 2002; Gay and Le Bas 2005; Harhoff et al. 1999; Rozhkov and Ivantcheva 1998; Trajtenberg 1990). Another study found a correlation between increases in company sales or profits and highly cited patents in the company's portfolio (Albert et al. 1991). Whether the backwards citations refer to the technological value, economic value, or both, there seems to be a correlation between the value or impact of a given patent and the backward citations it receives from future patents.

It has been noted that different types of inventive groups have different propensities to patent. These differences in patenting can lead to stratification of the data between technological areas, for example. As a result, it is often desirable to limit the data to a technological area in order to limit this stratification (Albert et al. 1991). In this study, the field of nanotechnology was chosen since it represents an emerging field that holds great promise in areas such as information technology, materials science, and medicine (Meyer 2001). Furthermore, experts believe that nanotechnology will be a key technology impacting almost every aspect of the economy (Hullmann and Meyer 2003; Meyer 2007) and the world market for nanotechnological products is forecasted to be as high as 150 billion dollars in 2010 and 2.6 trillion dollars in 2014 (Hullmann 2007).

While there is no consensus on the definition of nanotechnology (Meyer and Persson 1998), the definitions all relate to a field and collection of technologies (Hullmann 2007) which deal with the materials, structures, and physical phenomena that are in the physical size range of 1 nm (i.e., one billionth of a metre) to 100 nm as well as the techniques to study phenomena at these sizes (Braun et al. 1997; Kostoff et al. 2006). At this scale, the areas of physics and chemistry merge and novel properties of mater develop (Braun et al. 1997).

Several recent studies investigated information included in nanotechnology patents (Huang et al. 2004, 2006; Chen et al. 2007a, b) and nanotechnology literature (Kostoff et al. 2007), which motivates the research undertaken in this contribution. The goal of this study is to use machine learning techniques to examine information that can be derived from the front page of a patent in an attempt to better predict what impacts a patent's value. In other words, we want to determine (1) what information from within the patent has a bearing on future patent value (i.e., incoming citation counts) and (2) can a model be built to predict valuable patents based on the information contained in the patent?

To this end, classification and feature selection and ranking were performed to highlight the most significant features (attributes) from the derived information. Both, classification models and feature selection results show that the same three groupings of features, i.e., inventor and assignee performance and outgoing references, are relevant to determine feature patent value. The generated classification models are in the form of production rules and decision trees, which are easy to comprehend and apply to categorize future

patents. To the best of our knowledge, this paper is the first to apply machine learning methods to analyze patent data.

Methods

For this study, a representative, large set of patents was chosen based on the set of patent search terms described below. Then, using a custom crawler, information from the cover sheet of these patents was extracted and saved in a local database for processing. The data was then cleaned and processed using a selection of machine learning methods. First, the information from the cover page was converted into a set of numerical/nominal features, and next the feature values were fed into classification and feature selection algorithms to derive models and patterns that can be used to assess and predict the feature value of a given patent.

Data set

Since there is no unified, global patent database, data from the USPTO database was used for a number of reasons: (1) the USPTO database is the most representative because claims submitted in other countries are often simultaneously submitted in the US (Huang et al. 2003); (2) the US represents the largest commercial market in the world (Huang et al. 2003); (3) the US system is well developed with historical data in electronic format back to 1975 (Albert et al. 1991); and (4) the US system is the most universally representative system for analyzing international technology (Albert et al. 1991).

To retrieve the data, a list of nanotechnology related patents was created based on a set of search terms, which in turn were derived based on prior works that focused on analysis of nanotechnology patents. The early related contributions applied a simple approach by using only the keyword "nano*" to select relevant patents (Meyer 2001; Marinova and Mcaleer 2003; Hullmann and Meyer 2003). In one case, this was refined by excluding patents that included "nanosecond" and chemical compound "NaNO" keywords (Marinova and Mcaleer 2003). We applied an expanded set of keywords originally developed by researchers at the National Science Foundation (Huang et al. 2003), which was used in a number of recent contributions (Huang et al. 2003, 2004; 2006; Sampat 2004; Lee et al. 2006; Chen et al. 2007a, b). Similarly to the related studies (Meyer 2001; Hullmann and Meyer 2003; Lee et al. 2006), this keyword search was limited only to patent title and abstract to decrease the possibility of pulling out patents that may not pertain to nanotechnology, i.e., full text may mention nanotechnology only in passing. The keywords and the corresponding number of corresponding retrieved patents are summarized in Table 1. Similarly to (Lee et al. 2006), the records retrieved using these keywords were manually processed to remove irrelevant patents. More specifically, Lee and colleagues used the following keywords related to the "nano*" pattern to filter out irrelevant patents, "nanogram", "nanometer", "nanosecond", "nanoliter", "nanoampere", "nanofarad", "nanomole", the chemical compound "NaNO" and variations of those keywords. In our case the manual screening was based on an extended list of irrelevant keywords that include: "nanowatt", "nanosecond", "nano_second", "nano_sec", "microsecond", "millisecond", "nanogram", "microgram", "nano_mole", "nanomole", "nanonewton", "nano control store", "nanocode", "nanoROM", "nanoprogram", "nanokernel", "nano_kernel", "nanoperm", "nanodosimeter", "nanojoule", "nanogravity", "nanofarad", "nanoamp", "nano_amp", "nano_instruction", "nanomolar", "nano.sub" and

Table 1 Patent search terms

Search term	Patent count
Self-assembly*	32
Molecular electronics	528
Molecular motor	107
Molecular sensor	54
Quantum computing	192
Quantum dot(s)	1888
Quantum effect(s)	1219
(Self assembly) or (self assemble)	3547
(Atomic force microscopy) or (atomic force microscope) or (AFM)	7375
Atomic-force-microscope*	19
(Scanning tunnelling microscopy) or (scanning tunnelling microscope) or (STM)	5837
Scanning-tunnelling-microscope*	28
Atomistic simulation(s)	8
Biomotor	9
Molecular device(s)	4044
Nano* (later filtered to remove terms such as nanosecond, etc.)	119884
Total retrieved unique records	132670

Asterisk designates a wildcard and (s) designates a search for the singular and plural forms

variations of those keywords. We note that even if the applied search terms would produce some patents that do not concern nanotechnology, the machine learning methods applied to generate classification models, i.e., RIPPER and C4.5 classifiers, are tolerant to noise (Cohen 1995; Hilario and Kalousis 2000) and thus they can generate valid models in the presence of "noisy patents".

This list of patents was used as the seed for a custom crawler that extracted the cover sheet data from the USPTO online patent database; crawling was performed between June 8, 2007 and June 18, 2007. The resulting data was then cleaned to standardize naming conventions of inventors and assignees. First, the terms Incorporated, Limited, Corporation, LLC, etc. were standardized. For example, names containing "Incorporated", ", Incorporated", "Inc", ", Inc", or ", Inc. ", were changed to "Inc." for consistency. Then, "&" was replaced with "and" with the exception of "AT&T". A search was then done to find duplicate names when "Ltd. ", "Inc. ", "Co. ", etc. were removed. For example, ABC Inc. and ABC Ltd. Universities were manually cleaned to consolidate variations of "Governors of …", "Regents of …", "University of …", "President of …", etc. A list of duplicate names with different locations was then generated to search for and clean misspelled or inconsistent spellings of locations. Upon cleaning the inventor and assignee names, duplicate inventors and assignees were removed to provide a more accurate reflection of the inventor and assignee counts. Finally, based on the name, each assignee was assigned a type of Government, Company, Research, University, or Individual.

To assign a value to each patent, the number of incoming citations (i.e., backward citations) was used to determine a percentile for each patent by year. This percentile by year was used to normalize citation trends over time. Using this percentile, each patent was assigned one of two class labels: HIGH was assigned to any patents with a percentile greater than a threshold and denotes patent with high value; remaining patents were

categorized as LOW (with lower values). The threshold was set at 0.95 such that only patents in the top fifth percentile were assigned to the HIGH class. We also tested the developed models with threshold of 0.90 to investigate whether our conclusions depend on the threshold value.

While data was retrieved for patents ranging from 1976 to the present, not all features were reliable for all years. First, many features relied on information from previous years patents as discussed below while patents from the earlier years did not have all information. Furthermore, we want to ensure a reasonable number of incoming citations for each patent to reduce the amount of noise present in the first few years after patent publication (Hall et al. 2005). For example, Fig. 1 shows the average number of incoming citations and the number of patents per year for the extracted data set. Note that, while the number of patents continues to grow, the average number of incoming citations drops quickly after the turn of the century. As a result, the data for the last few years could not be used. Therefore, the period between 1990 and 2000 inclusive was chosen for the study to assure that the data is recent, and to allow sufficient data before and after the selected time span.

Feature representation

From the data retrieved by the crawler, 45 features were calculated to represent the information contained in each patent as shown in Table 5. Our aim was to use every piece of data from the patent's front page that could be converted in a quantifiable value (either numerical or nominal) in an automated fashion. Our comprehensive set of features was extracted based on different sections of the front page of each patent in the data set and encodes information about inventors, assignees, patent classification, and outgoing references. Most of these features display trends over time (e.g., year by year growth); therefore, they were normalized by dividing each value by the average of that value in each year. This enabled patents at the beginning of the data set period to be compared with those at the end of the data set period despite trends over time.

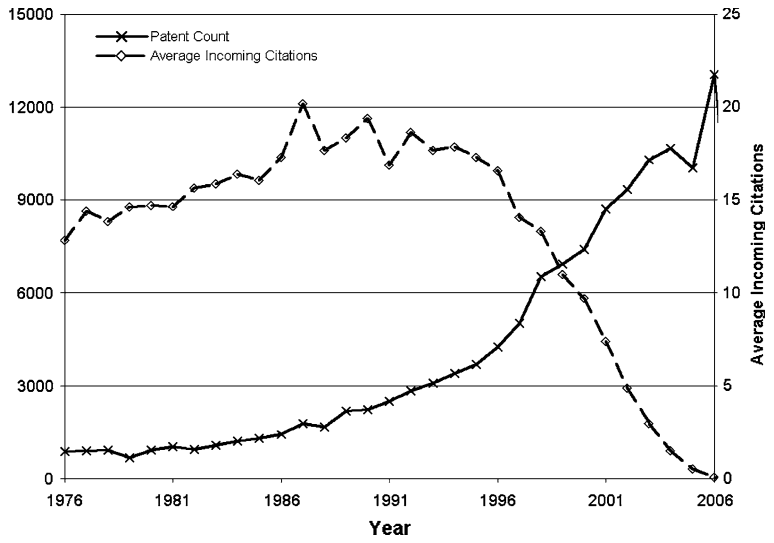

Fig. 1 Patent count and average incoming citations

Classifiers

A series of representative classification methods (classifiers) were used in an attempt to build a model representing the data set using the implementations found in Weka (Witten and Frank 2005). In other words, using the data found on the cover sheet of the patent and encoded into the corresponding 45 features, models were built to predict whether the patents would fall in the HIGH or LOW class. Such models can be used to predict the value (HIGH or LOW) for future patents, as well as to learn what patterns are characteristic for each of these two classification outcomes. The selected classifiers cover main types of classification methods including probabilistic methods (Naïve Bayes), regression (Logistic Regression), decision trees (C4.5 and Random Forest), and rule-based (RIPPER). We also applied meta cost wrapper in conjunction with the C4.5, RIPPER, and Random Forest classifiers in an attempt to improve the model performance. This extension allows focusing a given model to one of the predicted values (LOW or HIGH), i.e., a better quality model is build for one of the outcomes in expense of potentially poorer quality for the other outcome. Summary information for each of the applied classifiers is provided in Table 2. The Naïve Bayes and Regression classifiers generate models that are difficult to interpret (they consists of a set of numbers or equations) but they are computationally efficient (models can be generated quickly) and often provide good prediction quality. The remaining classifiers generate easy to interpret models (decision trees or IF…THEN… rules); these models can be analyzed to learn useful patterns that are associated with individual classification outcomes (LOW or HIGH patent value).

Evaluation of generated models

To evaluate models generated by each classifier, a 10-fold cross-validation test was used. In this test, the data was divided into ten randomly chosen subsets (folds). Then, 10 tests were performed with nine of the folds used as training data and a different fold used for testing during each test. The results of the 10 tests were then averaged to obtain the final results. This test procedure allows the reliable estimation of predictive quality of the developed models for future patent records.

A confusion matrix was generated showing the true positive (i.e., patents assigned to HIGH that are classified as HIGH), true negative (i.e., patents assigned to LOW that are classified as LOW), false positive (i.e., patents assigned to LOW but classified as HIGH), and false negative (i.e., patents assigned to HIGH but classified as LOW) for each classifier. The following four indices were computed to evaluate quality of each classification model: sensitivity, specificity, error rate reduction, and the ratio of true positive to false positive, see Table 3. Since the classifier's objective is to correctly classify instances (i.e., patents), it is desirable to maximize sensitivity, specificity, error rate reduction, and TP/FP ratio. However, many of these characteristics are optimized at the cost of others. For example, a highly sensitive classifier usually also has poor specificity.

Classifier optimization

Classifiers were optimized such that the sensitivity was kept >20% and the TP/FP ratio was maximized. Optimization involves adjustment of internal parameters of a given classifier in order to improve the quality of the generated models. Since the Naïve Bayes and Logistic classifiers were in general characterized by the lowest quality (see Results and discussion section) they were not optimized. At the same time, these two classifiers were kept to

Table 2 Summary of the applied classification methods

Name	Description	Model	Reference
Naïve Bayes	This method classifies a given patent by determining the probabilities of each class (HIGH/LOW) given the observed features according to Bayes' theorem. Given the known training data, the classifier calculates the probability that the observed patent is of HIGH value and that the observed patent is of LOW value. Then, for a given feature vector, it predicts the class with the greater probability	Set of conditional probabilities (difficult to interpret)	(John and Langley 1995)
Multinomial Logistic Regression	The multinomial Logistic Regression method with a ridge estimator was used. Using this method, given the observed features, the regression values for the HIGH and LOW classification outcomes are computed and the class with the larger value is chosen	Set of regression equations (difficult to interpret)	(Le Cessie and Van Houwelingen 1992)
C4.5	This method builds a decision tree in which each node represents a feature and each outgoing branch represents a test on that feature. When building the tree, at each node the feature with the highest value of information gain, a measure of the reduction in uncertainty of the result when the value of that feature is known, is selected. The observed features are matched against tree nodes, starting at the top of the tree, to provide the classification outcome (HIGH or LOW class)	Decision tree (easy to interpret)	(Quinlan 1993)
Random Forest	This method generates a set of decision trees, each using a randomly chosen subset of features. The combined set of trees forms a forest and each tree in the forest votes on the classification outcome.	Set of decision trees (easy to interpret)	(Breiman 2001)
RIPPER	RIPPER generates a set of rules by selecting features-value pairs (selectors) using the information gain. Next, it performs simplification by pruning individual selectors and entire rules in the rule set in order to better generalize the classification without increasing the error rate	Set of IF…THEN… rules (easy to interpret)	(Cohen 1995)
MetaCost	This method takes a base classifier (e.g., C4.5, RIPPER, etc.) and makes it cost sensitive such that the cost of false positive classifications (i.e., false HIGH patents) can differ from that of false negatives (i.e., false LOW patents). It does so by weighting the probability of each class by the cost and taking the lowest cost class for each instance (i.e., patent). This method was applied together with the C4.5, RIPPER, and Random Forest as the base classifiers	The same as the model of the base classifier.	(Domingos 1999)

Table 3 Summary of the computed quality indices

Definition	Description
Sensitivity $= \frac{TP}{TP+FN}$	Sensitivity represents the ratio of correctly identified positive instances. In our case, this represents the ratio of patents correctly identified as being in the HIGH class out of all the patents assigned to the HIGH class. This helps to determine the coverage of how many instances (patents) are correctly captured by the model
Specificity $= \frac{TN}{FP+TN}$	Specificity of a given classifier identifies the ratio of correctly identified negative instances. This represents the ratio of patents correctly identified as being in the LOW class out of all the patents assigned to the LOW class. This helps to determine how selective or exclusive the model is (i.e., how many patents are correctly excluded from the HIGH class)
Error rate $= \frac{FP+FN}{TP+TN+FP+FN}$	The error rate reduction represents the improvement between the baseline (i.e., assume all patents are in the LOW class) and the predictions of the model. The error rate is simply the ratio of incorrectly assigned instances out of all instances. Since our data consists of only two classes (i.e., HIGH and LOW), this is equivalent to the overall error rate of the classifier. Then, the error rate reduction is calculated by taking the difference between the error of the baseline and the error of the model. For example, if 5% of the instances are classified as HIGH, the error rate for the baseline of assuming all classes are LOW is 5%. If a given model has an error rate of 4%, this translates into a reduction in error rate of 20% (i.e. (5–4%)/5% = 20%)
Ratio of TP to FP $= \frac{TP}{FP}$	The ratio of true positive to false positive describes the quality of the resulting positive (HIGH) class. This describes the number of patents falsely classified as HIGH, which are in fact LOW, compared with the number of patents correctly classified as HIGH

TP true positive, *TN* true negative, *FP* false positive, *FN* false negative

provide a reference point for comparison with models generated by the remaining methods. The C4.5, RIPPER, Random Forest, and combinations of these with the MetaCost wrapper were optimized.

For the C4.5 classifier, two parameters were optimized: the confidence and the minimum number of instances per leaf. The confidence factor is used for pruning such that smaller confidence factors result in smaller trees. The other parameter is self-explanatory in that it imposes a minimum leaf size for the tree. For the RIPPER classifier, two parameters were optimized: the error pruning folds and the optimization runs. The error pruning folds control the amount of data that is used for growing the rules compared with the data used for pruning the rules. One fold is always used for pruning data while the remaining folds are used for growing the rules. The optimization runs is also self-explanatory in that it controls how many runs are performed for optimizing the RIPPER algorithm. For the Random Forest, the number of trees generated for the forest was optimized. For the MetaCost classifier, the ratio of false negative cost to false positive cost was optimized for the C4.5, RIPPER, and Random Forest classifiers.

To speed up the optimization, only threefold cross-validation was used. This provides quicker results since fewer tests were performed. Furthermore, since the randomness of the folds impacts the performance of the classifiers, a minimum of five repetitions were performed on the top candidate combinations of parameters. In other words, the tests with combinations of parameters that performed best on each classifier were repeated to ensure that the better performance was not the result of an optimal selection of random folds for that classifier.

Feature selection

Three methods were used for feature selection: the χ^2 method, the Gain Ratio method (Quinlan 1993), and the ReliefF method (Kononenko 1994). The χ^2 method ranks the features' relevance based on their χ^2 statistic. The Gain Ratio method ranks the features by measuring the gain ratio of each feature with respect to the class. Finally, the ReliefF method ranks the features by sampling patents then evaluating the value of each feature for the nearest patent of the same and different classes. The implementations found in Weka (Witten and Frank 2005) for all three methods were used within the regime of the 10-fold cross-validation. The ranking of features provided by each of the three methods was averaged over the 10-folds to find the most significant features.

Results and discussion

First, the models generated by the considered classifiers are tested to verify whether they could be successfully used to predict and find factors related to the patent's value. Next, a feature selection study was conducted to find the most promising features that can be used to differentiate between HIGH and LOW value patents. Finally, selected, best performing models, i.e., decision tree generated by C4.5 and set of IF…THEN rules generated by RIPPER, are presented and analyzed in the context of the feature selection results.

Classifier testing

The eight classifiers (i.e., Naïve Bayes, Logistic Regression, C4.5, RIPPER, Random Forest, MetaCost with C4.5, MetaCost with RIPPER, and MetaCost with Random Forest) were tested on two data sets. The first was the data set with the HIGH class defined as the top fifth percentile (per year). The second case was the same data but with the HIGH class defined as the top tenth percentile. This contrasted the performance with two different sizes of HIGH classes (i.e., 10% of the patents versus 5% of the patents). Tenfold cross validation tests were performed and reported for both data sets. This testing was to (1) compare the performance of the classifiers; and (2) to compare evaluation with two different thresholds for the HIGH class definition. The quality of the generated classification models and the optimized values of the classifier parameters are summarized in Table 4.

In all three cases, the Naïve Bayes classification performed worse than the default of classifying all patents as LOW. This is shown by the negative error rate reduction meaning that the error rate of using the Naïve Bayes classifier is greater (i.e., performs worse) than assuming all patents are classified as LOW. Also, in all three cases, the six optimized classifiers (i.e., C4.5, RIPPER, Random Forest, MetaCost C4.5, MetaCost RIPPER, and MetaCost Random Forest) performed significantly better than the Logistic Regression and Naïve Bayes in terms of TP/FP ratio, sensitivity, and error rate reduction. For example, the RIPPER classifier improved by 36% in TP/FP, 124% in sensitivity, and 226% in error rate reduction relative to the Logistic Regression. At the same time, there was little change in the specificity from the Logistic Regression to the optimized classifiers.

When dropping the class threshold from the top fifth percentile to the top tenth percentile, there was only a small change in the performance of the classifiers suggesting that the models are insensitive to changes in this threshold.

In all three cases, the MetaCost wrapper improved the TP/FP ratio and lowered the error rate reduction and sensitivity relative to the base classifier. In other words, if we consider

Table 4 Performance and optimized parameter values for the considered classifiers

Data set	Quality index	Naïve Bayes	Logistic Regression	C4.5	RIPPER	Random Forest	MetaCost w/ C4.5	MetaCost w/ RIPPER	MetaCost w/ Random Forest
Top fifth percentile	TP/FP ratio	0.15	1.54	2.86	2.09	2.75	3.54	3.18	2.82
	Sensitivity (%)	41.6	11.9	24.1	26.6	21.5	21.1	21.2	20.7
	Specificity (%)	84.9	99.6	99.6	99.3	99.6	99.7	99.6	99.6
	Error rate reduction (%)	−238	4.2	15.6	13.7	13.6	15.1	14.5	13.3
Top tenth percentile	TP/FP Ratio	0.27	1.74	3.75	2.48	2.72	4.30	3.18	2.79
	Sensitivity (%)	17.7	12.0	22.7	28.1	24.5	20.6	25.4	24.6
	Specificity (%)	92.7	99.2	99.3	98.7	99.0	99.4	99.0	99.0
	Error rate reduction (%)	−47	5.1	16.6	16.7	15.5	15.9	16.6	15.8

Optimized values of the classifier parameters	
C4.5	Confidence: 0.025, minimum instances per leaf: 6
RIPPER	Error pruning folds: 2, optimization runs 3
Random Forest	12 Trees
MetaCost with C4.5	Confidence: 0.025, minimum instances per leaf: 6, cost matrix: $\begin{bmatrix} 0 & 1 \\ 1.5 & 0 \end{bmatrix}$
MetaCost with RIPPER	Error pruning folds: 2, optimization runs 3, cost matrix: $\begin{bmatrix} 0 & 1 \\ 1.25 & 0 \end{bmatrix}$
MetaCost with Random Forest	12 Trees, cost matrix: $\begin{bmatrix} 0 & 1.25 \\ 1 & 0 \end{bmatrix}$

the classifier with the highest TP/FP ratio for both data sets, the MetaCost with C4.5 classifier, for every three correctly classified HIGH instances, there will be one incorrectly classified HIGH instance (i.e., actually LOW). Meanwhile, with the C4.5 classifier, for every 3.5 correctly classified HIGH instances, there will be one instance incorrectly classified as HIGH. This is an improvement of $\sim$20%. However, with this increased TP/FP ratio comes a decreased error rate reduction. For example, with the MetaCost with C4.5 classifier compared with the C4.5 classifier, the error rate reduction falls from 15.6% without the MetaCost wrapper to 15.1% with the wrapper. In other words, the error rate reduction is 3% lower when using the MetaCost wrapper. The sensitivity between the C4.5 and MetaCost with C4.5 classifiers decreased by 12% implying that the MetaCost wrapper covers fewer HIGH instances. It should also be noted that the MetaCost wrapper did not impact the error rate reduction for the RIPPER classifier with the top tenth percentile data and even improved the error rate reduction in the top fifth percentile case. However, the sensitivity still fell by 20% when using the MetaCost wrapper over the ordinary RIPPER classifier.

Overall, the C4.5 and RIPPER classifiers have the best error rate reduction of all classifiers tested, and are characterized by over 20% sensitivity (they correctly predict over 20% of the HIGH value patents) and very high, over 99% specificity (they correctly exclude over 99% of the LOW value patents). For the top fifth percentile tests, the C4.5 performed best while with the top tenth percentile tests, the RIPPER performed marginally better in terms of the error rate reduction. Meanwhile, the C4.5 classifier (with MetaCost) performed best in terms of TP/FP ratio for both tests (i.e., top fifth and top tenth percentiles). The results for the C4.5 and RIPPER classifiers indicate that the corresponding models can be used to successfully predict whether a given patent has the potential to be valuable.

Feature selection

The results of the feature selection using three methods (χ^2, Gain Ratio, and ReliefF) are shown in Table 5 along with the rank based on the averages of the three methods. The average rank is not an average of the other three rank values but a rank based on the average. For example, the feature entitled 'in_topInventors' has ranks of 1, 1, and 32 for the χ^2, Gain Ratio, and ReliefF, respectively. While the average of these ranks is 10.7, there are only two features with rank averages <10.7, namely count_assig_refs with 4.0 and avg_inventors_refs 9.0. As a result, this feature is ranked as number 3, not 10.7.

The χ^2 and Gain Ratio methods resulted in relatively similar rankings of features. However, the ReliefF method provided different results. This difference could suggest the presence of noise in the data since the ReliefF considers specific neighbouring data points while the other two methods look at the classes in their entirety.

The average rank relative to the overall rank is shown in Fig. 2 along with the growth rate of this average rank. The figure shows alternating regions of relatively quick growth in the average rank and relatively slow growth in the relative rank. The growth begins with a quick rise in the average rank until the fourth feature followed by a period of slow growth until a spike in the growth for features ranked 11, 12, and 13 overall. This is followed by another slow growth period before the remaining period of relatively quick growth after rank 23. Note that there is a gap in the plot in Fig. 2 at an overall rank of 15 since there are two features tied for a rank of 14. There is another gap from 24 to 26 since there are four features tied with a rank of 23. These alternating regions of high growth followed by slow growth imply natural groupings of feature relevance suggesting a natural break for feature

Table 5 Feature definitions and results of the feature selection

Group	Feature	Description	Ranks			
			χ^2	Gain Ratio	ReliefF	Overall
Outgoing references	**count_out_refs**	Number of distinct outgoing patent references listed on the front page	7	9	22	4
	count_out_sameinv	Count of distinct outgoing patent references with at least one inventor in common	29	33	32	35
	count_out_diffinv	*count_out_refs − count_out_sameinv*	7	12	24	8
	count_out_sameUSclass	Count of distinct outgoing patent references with at least one US class in common	10	16	29	15
	count_out_same1stUSclass	Count of distinct outgoing patent references with the first US class in common	19	21	25	24
	count_out_diffUSclass	Count of distinct outgoing patent references with no US classes in common	9	14	21	9
	count_out_regions	Count of distinct regions based on the inventors of outgoing patent references	28	19	11	19
	avg_out_inc_refs	Avg/min/max incoming reference count for the outgoing patent references	13	11	16	5
	min_out_inc_refs		21	26	18	24
	max_out_inc_refs		12	18	26	17
	pat_refage_avg	Avg/min/max age of referenced patents	15	20	35	28
	pat_refage_min		17	24	42	31
	pat_refage_max		27	13	37	29
Inventors	count_inventors	Count of inventors assigned to the patent	22	22	13	18
	avg_inventors_refs	Avg/min/max across all inventors of the patent of the ratio of incoming citations to patents of each inventor divided by the average references per inventor for the year of the patent	4	6	20	2
	min_inventors_refs		6	10	37	14
	max_inventors_refs		5	8	27	5
	avg_inventors_pats	Avg/min/max across all inventors of the patent of the ratio of patents of each inventor divided by the average references per inventor for the year of the patent	15	29	17	22
	min_inventors_pats		18	23	18	20
	max_inventors_pats		11	25	9	11

Table 5 continued

Group	Feature	Description	Ranks			
			χ^2	Gain Ratio	ReliefF	Overall
	avg_inventors_invent_time	Avg/min/max across all inventors of the patent of the difference between the current patent and the earliest patent for the inventor	32	7	11	13
	min_inventors_invent_time		44	45	1	34
	max_inventors_invent_time		40	42	14	36
	ratio_inventor_samelocation	MAX(inventors in each location) ÷ $COUNT$(*inventors*)	26	32	7	24
	ratio_inventor_sameregion	MAX(inventors in each region) ÷ $COUNT$(*inventors*)	25	31	9	24
	in_topInventors	True if one of the inventors has an average incoming citation count greater than the 95th percentile	1	1	33	3
	avg_inventor_lastpatelapsed	Avg/min/max time since the last patent of the inventors was published	30	34	45	41
	min_inventor_lastpatelapsed		40	43	3	33
	max_inventor_lastpatelapsed		43	41	1	32
Classes	in_topClasses	True if the patent is assigned to a class that has an average incoming citation count greater than the 95th percentile	20	3	37	21
	in_1stTopClasses	True if the patent's first assigned class has an average incoming citation count (limited to patents with this class assigned as the first class) greater than the 95th percentile	35	5	15	15
	count_classes_int	Count of international classes assigned to the patent	22	15	4	7
	count_classes_US	Count of US classes assigned to the patent	24	17	5	12
Assignee	**in_topAssignees**	True if the assignee has an average incoming citation count greater than the 95th percentile	2	2	40	9
	assig_type	Type of assignee among Company, Government, University, Research, Individual, or Invalid	33	35	31	38
	assig_corp	True if assig_type is Company	31	36	30	37
	assig_govt	True if assig_type is Government	38	28	34	39
	assig_indiv	True if assig_type is Individual	45	44	41	45
	assig_research	True if assig_type is Research	34	29	44	40
	assig_univ	True if assig_type is University	39	38	43	44
	count_assig_refs	Count of incoming references for the assignee (limited to the training period) normalized by the average assignee references for the year of the patent	3	4	8	1

Table 5 continued

Group	Feature	Description	Ranks			
			χ^2	Gain Ratio	ReliefF	Overall
	count_assig_pats	Count of patents for the assignee	36	37	6	30
	assig_lastpatelapsed	Time since the last patent of the assignee was published	37	39	35	43
	count_nonpat_refs	Count of non-patent references	14	27	23	23
	diff_filedpub	Difference between the field date and the published date	42	40	28	42

Notes: All features with varying averages by year were normalized per year (i.e., divided by the average value for that year). The top 14 features are denoted in bold

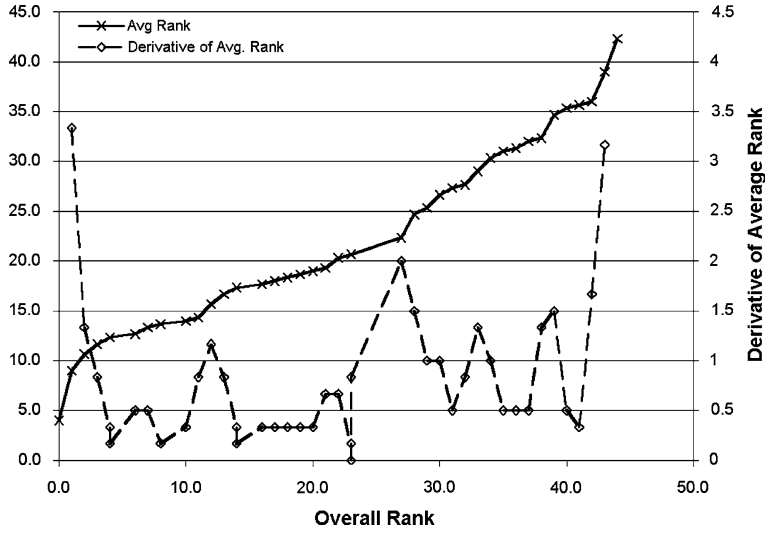

Fig. 2 Feature selection ranks and growth rate

selection at rank 14; the corresponding features are shown in bold in Table 5. Given the quick growth at an overall rank between 27 and 29, this implies another natural break for feature selection before the last group of features.

These rankings suggest several factors that impact the value of a patent. In general, we observe that features describing some aspects all four major categories of information that can be extracted from the patent's front page, i.e., inventors, assignees, patent classification, and outgoing references, were found important. First, the performance of the inventors and assignees seem to be important. For example, assignees and inventors that already have valuable patents tend to impact the value of their future patents (e.g., fields count_assig_refs, avg_inventors_refs, in_topInventors, max_inventors_refs, in_topAssignees, min_inventors_refs). Also, inventors with fewer patents and who have been inventing longer, seem to the value of future patents (e.g., fields max_inventors_pats, avg_inventors_invent_time). The outgoing references or the science that the invention is based upon also seem to have an impact such as the performance of referenced patents and the count of patents both in total and from different inventors or classes (e.g., fields count_out_refs, avg_out_inc_refs, count_out_diffinv, count_out_diffUSclass). Finally, the number of international and US classes assigned to the patent also seem to impact the value of the patent (e.g., fields count_classes_int, count_classes_US).

Rule model

The rule-based model shown in Table 6 was generated by the RIPPER classifier. The top five features from the feature selection are shown in bold and appear frequently throughout the rules suggesting agreement between the two methods. If we consider rules 2, 7, and 9 as examples since they describe a relatively high number of correct predictions and a low number of incorrect predictions (relative to the number of correct) when these rules would be used on our data set, they suggest the following. First, patents from top performing inventors (i.e., in_topInventors $\geq$ 1) who have also filed a low total

Table 6 Rule based model generated by RIPPER classifier

Rule no.	Antecedents/conditions	Class	No. of predictions (correct/incorrect)
1	(**in_topInventors** >= 1) and (min_inventors_invent_time >= 0.483741) and (**count_out_refs** >= 0.021341) and (pat_refage_avg <= 0.616153) THEN	class=HIGH	(93.0/15.0)
2	(**in_topInventors** >= 1) and (max_inventors_pats <= 0.033291) THEN	class=HIGH	(160.0/52.0)
3	(**in_topInventors** >= 1) and (**count_assig_refs** >= 0.010193) and (assig_lastpatelapsed >= 0.545192) THEN	class=HIGH	(153.0/61.0)
4	(**in_topInventors** >= 1) and (count_out_diffinv >= 0.014126) and (min_inventors_invent_time >= 0.42845) and (count_assig_refs >= 0.009475) and (min_inventors_refs >= 0.019056) THEN	class=HIGH	(37.0/8.0)
5	(avg_inventors_refs >= 0.009188) and (in_topInventors >= 1) and (max_inventors_invent_time >= 0.482573) and (**avg_out_inc_refs** >= 0.043995) and (**avg_out_inc_refs** >= 0.058027) and (**avg_out_inc_refs** <= 0.083892) THEN	class=HIGH	(29.0/4.0)
6	(**max_inventors_refs** >= 0.008154) and (**in_topInventors** >= 1) and (count_out_diffinv >= 0.020207) and (max_inventors_pats <= 0.079062) and (max_inventors_pats >= 0.05628) THEN	class=HIGH	(32.0/10.0)
7	(**count_assig_refs** >= 0.001138) and (count_assig_pats <= 0.001446) and (count_assig_pats <= 0.000739) THEN	class=HIGH	(214.0/5.0)
8	(**max_inventors_refs** >= 0.008346) and (**in_topInventors** >= 1) and (max_inventors_invent_time >= 0.484507) and (count_assig_refs >= 0.010498) and (max_inventors_pats >= 0.125315) THEN	class=HIGH	(67.0/27.0)
9	(**max_inventors_refs** >= 0.008135) and (avg_inventors_invent_time >= 0.467675) and (**count_assig_refs** >= 0.001803) and (count_assig_pats <= 0.00467) and (count_assig_pats <= 0.002438) and (count_assig_pats <= 0.00126) and (pat_refage_min <= 0.996644) THEN	class=HIGH	(43.0/7.0)
10	(**count_assig_refs** >= 0.001698) and (count_assig_pats <= 0.004986) and (count_assig_pats <= 0.002054) and (count_assig_pats <= 0.001259) and (avg_inventor_lastpatelapsed <= 0) and (min_out_inc_refs <= 0.005831) THEN	class=HIGH	(34.0/9.0)
11	(**count_assig_refs** >= 0.001716) and (count_assig_pats <= 0.004783) and (count_assig_pats <= 0.001878) and (count_classes_int >= 0.139949) and (min_inventors_pats >= 0.01949) and (min_inventors_pats <= 0.031535) THEN	class=HIGH	(30.0/10.0)
12	(**max_inventors_refs** >= 0.007298) and (**count_assig_refs** >= 0.003051) and (count_assig_pats <= 0.007115) and (count_assig_pats <= 0.00377) and (min_inventors_refs >= 0.026782) and (diff_filedpub >= 0.239321) and (max_inventors_invent_time >= 0.456638) THEN	class=HIGH	(26.0/7.0)
13	(**avg_inventors_refs** >= 0.010249) and (avg_inventors_invent_time >= 0.469245) and (**avg_inventors_refs** >= 0.051114) and (avg_inventors_invent_time >= 0.487538) and (max_inventors_pats >= 0.165754) and (min_out_inc_refs >= 0.008192) THEN	class=HIGH	(63.0/24.0)
14	(**count_assig_refs** >= 0.00303) and (count_assig_pats <= 0.011979) and (count_assig_pats <= 0.003451) and (diff_filedpub >= 0.262923) and (pat_refage_max <= 0.075085) THEN	class=HIGH	(23.0/8.0)

Table 6 continued

Rule no.	Antecedents/conditions	Class	No. of predictions (correct/incorrect)
15	(**count_assig_refs** >= 0.004373) and (count_assig_pats <= 0.011893) and (count_assig_pats <= 0.005143) and (count_assig_pats <= 0.003116) and (avg_inventors_invent_time >= 0.492209) and (**avg_inventors_refs** >= 0.006738) THEN	class=HIGH	(30.0/10.0)
16	ELSE	class=LOW	(46847.0/1667.0)

Notes: Model parameters. Class threshold of top fifth percentile, three optimization runs, two error pruning folds. Bolded features are in top five of the feature selection ranking. The right most column shows the number of correct/incorrect predictions when these rules would be used on our data set

number of patents (i.e., max_inventors_pats $\leq$ 0.03) perform relatively well. This also suggests that there may be a number of inventors who have many low value patents but only one or two high value patents (i.e., those who appear in the top inventors list but who have max_inventors_pats $>$ 0.03). Similarly, rule seven suggests that assignees who are referenced (i.e., count_assig_refs $\geq$ 0.001) but who have a low total number of patents (i.e., count_assig_pats $\leq$ 0.0007) produce high valued patents. Rule nine suggests that inventors who are well referenced (i.e., max_inventors_refs $\geq$ 0.008), who are not new to inventing (i.e., avg_inventors_invent_time $\geq$ 0.47), who are associated with assignees that are referenced (i.e., count_assig_refs $\geq$ 0.002) but don't have too many patents (i.e., count_assig_pats $\leq$ 0.001) are likely to produce high valued patents. Finally, rules 1, 4 and 6 suggest that higher count of outgoing references is also associated with more valuable patents. This concerns both the total count (i.e., count_out_refs $\geq$ 0.021341 in rule 1) and the count of references that were not developed by the patent's authors (i.e., count_out_diffinv $\geq$ 0.014126 in rule 4 and count_out_diffinv $\geq$ 0.020207 in rule 6).

Overall, the RIPPER's model shows that valuable patents are associated with inventors who were already successful in filling valuable patents and have a history of inventing, assignees that were referenced and that filed relatively low number of patents, and have a higher count of references.

Decision tree model

The decision tree shown in Table 7 was generated by the C4.5 algorithm. The test at the top of the tree is listed at the top of the left column in the table, i.e., in_topInventors $\leq$ 0 and in_topInventors $>$ 0, and the corresponding branches are denoted by tests at positions which are connected by vertical lines. Tree can be converted into a set of rules by following the tests (nodes) connected by branches from the root node (the top most node) to leaves (bottom most nodes). The leaves are denoted by their corresponding classification outcomes values (LOW or HIGH) and proportion of correct/incorrect prediction when the corresponding rule would be used to classify patents from our data set. The top five features from the feature selection are shown in bold in Table 7 and appear throughout the tree, especially in the first few levels (e.g., in_topInventors, count_assig_refs), which shows agreement between this model and the results of our feature selection. Note that, the decision tree shows leaves (and the corresponding rules) for both HIGH and LOW classes, while the RIPPER's rules only represent the HIGH classes with all instances not fitting within the rules assigned to LOW class.

While the rules that can be derived from the decision tree vary from the RIPPER rules, certain trends remain. For example, instances that do not have inventors on the top inventors list (i.e., in_topInventors above 0) tend to have LOW value. This is amplified when combined with a low number of assignee references (i.e., count_assig_refs). For example, of the instances with in_topInventors $\leq$ 0 in the C4.5 model, 96% are of LOW value. If the assignee references are also low (i.e., count_assig_refs $<$ 0.002), this increases to 98% with LOW value. The corresponding rule (based on the top most leaf in the tree shown in Table 7) reads "IF in_topInventors $\leq$ 0 AND count_assig_refs $\leq$ 0.00174 AND count_assig_pats $\leq$ 0.000518 AND count_assig_refs $\leq$ 0.001009 THEN LOW". It shows that patents with inventors who did not produce valuable patents so far, and which are assigned to institutions with low count of patents and references are likely to be of low value, i.e., in 100%*8316/(8316 + 193) = 98% of cases they will be of LOW value. Rules 1 and 3 from Table 6 indirectly imply this relation as well since they show a trend towards HIGH patent

Table 7 Decision tree model generated by C4.5 classifier

in_topInventors <= 0
| **count_assig_refs** <= 0.00174
| | count_assig_pats <= 0.000518
| | | **count_assig_refs** <= 0.001009: LOW (8316.0/193.0)
| | | **count_assig_refs** > 0.001009
| | | | **count_assig_refs** <= 0.00111
| | | | | avg_inventors_invent_time <= 0.507175: HIGH (26.0/4.0)
| | | | | avg_inventors_invent_time > 0.507175: LOW (22.0)
| | | | **count_assig_refs** > 0.00111: HIGH (144.0/5.0)
| | count_assig_pats > 0.000518: LOW (8679.0/58.0)
| **count_assig_refs** > 0.00174
| | count_assig_pats <= 0.001466
| | | count_assig_pats <= 0.000866: HIGH (98.0/14.0)
| | | count_assig_pats > 0.000866: LOW (221.0/93.0)
| | count_assig_pats > 0.001466: LOW (29333.0/1341.0)
in_topInventors > 0
| min_inventors_invent_time <= 0.514067
| | pat_refage_avg <= 0.607189: LOW (26.08/1.08)
| | pat_refage_avg > 0.607189
| | | **count_assig_refs** <= 0.001629: LOW (88.0/16.0)
| | | **count_assig_refs** > 0.001629
| | | | count_assig_pats <= 0.001901
| | | | | max_inventors_pats <= 0.136219: HIGH (75.97/6.0)
| | | | | max_inventors_pats > 0.136219: LOW (8.0/3.0)
| | | | count_assig_pats > 0.001901
| | | | | **count_assig_refs** <= 0.005072: LOW (66.0/4.0)
| | | | | **count_assig_refs** > 0.005072
| | | | | | pat_refage_min <= 0.996726
| | | | | | | pat_refage_avg <= 0.617823
| | | | | | | | **max_inventors_refs** <= 0.044886
| | | | | | | | | pat_refage_max <= 0.076603
| | | | | | | | | | pat_refage_avg <= 0.609137
| | | | | | | | | | | ratio_inventor_samelocation <= 0.029905: HIGH (20.08/2.0)
| | | | | | | | | | | ratio_inventor_samelocation > 0.029905
| | | | | | | | | | | | **avg_inventors_refs** <= 0.016384: LOW (7.0/1.0)
| | | | | | | | | | | | **avg_inventors_refs** > 0.016384: HIGH (6.08/2.0)
| | | | | | | | | pat_refage_avg > 0.609137
| | | | | | | | | | count_inventors <= 0.126454
| | | | | | | | | | | count_out_regions <= 0.125
| | | | | | | | | | | | diff_filedpub <= 0.215571: LOW (22.0/1.0)
| | | | | | | | | | | | diff_filedpub > 0.215571
| | | | | | | | | | | | | count_nonpat_refs <= 0.001936
| | | | | | | | | | | | | | count_out_diffinv <= 0.011641
| | | | | | | | | | | | | | | **avg_out_inc_refs** <= 0.021854: LOW (20.63/4.63)
| | | | | | | | | | | | | | | **avg_out_inc_refs** > 0.021854: HIGH (9.0)
| | | | | | | | | | | | | | count_out_diffinv > 0.011641: LOW (18.0)
| | | | | | | | | | | | | count_nonpat_refs > 0.001936
| | | | | | | | | | | | | | **count_assig_refs** <= 0.122026
| | | | | | | | | | | | | | | count_out_sameinv <= 0
| | | | | | | | | | | | | | | | **max_inventors_refs** <= 0.024698: HIGH (28.0/9.0)
| | | | | | | | | | | | | | | | **max_inventors_refs** > 0.024698: LOW (16.0/4.0)
| | | | | | | | | | | | | | | count_out_sameinv > 0: LOW (23.0/5.0)
| | | | | | | | | | | | | | | **count_assig_refs** > 0.122026: HIGH (22.0/4.0)

…continued from the left-hand-side column

| | | | | | | | | | | | | count_out_regions > 0.125
| | | | | | | | | | | | | | **count_assig_refs** <= 0.261245
| | | | | | | | | | | | | | | min_inventors_pats <= 0.015318: LOW (6.0)
| | | | | | | | | | | | | | | min_inventors_pats > 0.015318: HIGH (6.0/2.0)
| | | | | | | | | | | | | | **count_assig_refs** > 0.261245: HIGH (9.0)
| | | | | | | | | | | | | count_inventors > 0.126454
| | | | | | | | | | | | | | count_out_diffUSclass <= 0.030924: LOW (64.63/4.63)
| | | | | | | | | | | | | | count_out_diffUSclass > 0.030924: HIGH (6.0/2.0)
| | | | | | | | | pat_refage_max > 0.076603
| | | | | | | | | | count_out_sameUSclass <= 0.052881: HIGH (36.23/6.0)
| | | | | | | | | | count_out_sameUSclass > 0.052881: LOW (9.0/1.0)
| | | | | | | | **max_inventors_refs** > 0.044886
| | | | | | | | | count_out_same1stUSclass <= 0.12441
| | | | | | | | | | max_inventors_invent_time <= 0.482632
| | | | | | | | | | | assig_type = Company
| | | | | | | | | | | | **count_assig_refs** <= 0.008184: LOW (8.0)
| | | | | | | | | | | | **count_assig_refs** > 0.008184
| | | | | | | | | | | | | count_assig_pats <= 0.27064
| | | | | | | | | | | | | | **count_assig_refs** <= 0.047848
| | | | | | | | | | | | | | | avg_inventor_lastpatelapsed <= 0.007198: LOW (7.0)
| | | | | | | | | | | | | | | avg_inventor_lastpatelapsed > 0.007198
| | | | | | | | | | | | | | | | max_inventors_invent_time <= 0.420221: LOW (9.0/1.0)
| | | | | | | | | | | | | | | | max_inventors_invent_time > 0.420221: HIGH (23.0/5.0)
| | | | | | | | | | | | | | **count_assig_refs** > 0.047848: HIGH (9.0)
| | | | | | | | | | | | | count_assig_pats > 0.27064: LOW (8.0)
| | | | | | | | | | assig_type = Government: HIGH (0.0)
| | | | | | | | | | assig_type = University: HIGH (16.0/3.0)
| | | | | | | | | | assig_type = Research: LOW (1.0)
| | | | | | | | | | assig_type = Individual: HIGH (1.0)
| | | | | | | | | | assig_type = Invalid: HIGH (0.0)
| | | | | | | | | max_inventors_invent_time > 0.482632: HIGH (135.0/33.0)
| | | | | | | | count_out_same1stUSclass > 0.12441: LOW (21.0/4.0)
| | | | | | | pat_refage_avg > 0.617823
| | | | | | | | max_inventors_invent_time <= 0.425354: LOW (20.1/0.1)
| | | | | | | | max_inventors_invent_time > 0.425354
| | | | | | | | | avg_inventor_lastpatelapsed <= 0.015069: LOW (8.0)
| | | | | | | | | avg_inventor_lastpatelapsed > 0.015069
| | | | | | | | | max_inventor_lastpatelapsed <= 0.040922: HIGH (12.0)
| | | | | | | | | max_inventor_lastpatelapsed > 0.040922: LOW (31.1/5.1)
| | | | | | pat_refage_min > 0.996726: HIGH (25.07/5.0)
| min_inventors_invent_time > 0.514067
| | count_assig_pats <= 0.008088: HIGH (61.0)
| | count_assig_pats > 0.008088
| | | **count_assig_refs** <= 0.005989: LOW (6.0)
| | | **count_assig_refs** > 0.005989
| | | | avg_inventor_lastpatelapsed <= 0.009083: HIGH (41.0/6.0)
| | | | avg_inventor_lastpatelapsed > 0.009083: LOW (7.0/2.0)

Notes: Model parameters. Confidence of 0.025 and a minimum of six instances per leaf. Bolded features are in top five of the feature selection ranking. Numbers in brackets located at the leaf nodes show the number of correct/incorrect predictions when the corresponding rules would be used on our data set

value from patents with inventors on the top inventors list and assignees that are well referenced.

We also analyze two sample strong rules that are associated with patents of HIGH value. The first rule reads "IF in_topInventors $>$ 0 AND min_inventors_invent_time $>$ 0.514067 AND count_assig_pats $\leq$ 0.008088 THEN HIGH". This rule is consistent with the model generated by RIPPER (e.g. rules 1 and 7) and shows that valuable patents are associated with inventors who already filed valuable patents (i.e., in_topInventors $>$ 0) and who have above average experience in filing patents (i.e., they were inventing patents for a long time; min_inventors_invent_time $>$ 0.514067), and with assignees which filed relatively low number of patents (i.e., count_assig_pats $\leq$ 0.008088). This rule describes 61 patents of HIGH value from our data set, while excluding all patents of LOW value. The second rule "IF in_topInventors $>$ 0 AND min_inventors_invent_time $\leq$ 0.514067 AND pat_refage_avg $>$ 0.607189 AND count_assig_refs $>$ 0.001629 AND max_inventors_pats $\leq$ 0.136219 THEN HIGH" shows that inventors who have history of filing valuable patents (i.e., in_topInventors $>$ 0), filed a low total number of patents (i.e., max_inventors_pats $\leq$ 0.136219) and who invent for a shorter than average time (i.e., min_inventors_invent_time $\leq$ 0.514067) can also file valuable

patents. According to this rule, patents invented by these individuals would have to reference patents that have above average age (i.e., were published relatively long time ago; pat_refage_avg > 0.607189) and would have assignees who are referenced (i.e., count_assig_refs > 0.001629) to be of HIGH value. The second rule is correct in $100\%*76/(76 + 6) = 93\%$ of cases for the patents in our data set.

In general, for the HIGH class the decision tree shows patterns that are consistent with the RIPPER's model, i.e., valuable patents are associated with inventors who already filed valuable patents, and with assignees who filed relatively low number of patents and who are referenced. At the same time, the tree shows that some of the patents of LOW value are filed by inventors who did not invent highly valuable patents yet, and are assigned to institutions (assignees) with low count of filed patents and references to their patents.

Conclusions

Since patents represent the technological change or inventive activity and output, using patinformatics and patent bibliometrics can help take information from patents to target innovative efforts. For example, by considering the top factors impacting patent success, efforts can be focused on patents with high probabilities of success.

We used an array of machine learning techniques to find the most informative factors (expressed as numerical features) that allow differentiating between patents of HIGH and LOW value within the field of nanotechnology. We encoded a large set nanotechnology patents using a predefined set of features to develop, test and analyze several optimized classification models. The six optimized classifiers tested (i.e., C4.5, RIPPER, Random Forest, MetaCost C4.5, MetaCost RIPPER, & MetaCost Random Forest) as well as Logistic Regression performed better than the default of assigning all patents to the LOW class. This implies that a pattern to patent success exists. The top performing C4.5 and RIPPER classification models achieved over 20% sensitivity (i.e., they covered above 20% of the valuable patents), over 99% specificity (i.e., they succeeded in excluding over 99% of the LOW valued patents), and were characterized by TP/FP ratio that shows that for every three correctly classified HIGH patents, only one incorrect classification would be performed. Furthermore, the relatively unchanged performance when varying the HIGH class threshold from the top fifth percentile to the top tenth percentile suggests that this pattern may exist regardless of reasonable changes in this threshold.

The classification models and feature selection results highlight consistent trends in the data that can be used to target efforts towards patents in nanotechnology that are more likely to perform well. The following factors were identified as being associated with the value of the patent:

- Patents originating from inventors that have already performed well (i.e., who published valuable patents and have been inventing longer) seem to have a higher probability of performing well.
- Some of the valuable patents are assigned to institutions or individuals who were referenced before and which filed relatively low number of patents.
- Patents with inventors who did not invent valuable patents yet and which are assigned to institutions (assignees) with low count of filed patents and references to their patents seem to have a higher probability of being of low value.

– The count of outgoing references, which denote the scientific sources for a given invention, is also shown to impact the value of the patent, i.e., higher count of outgoing references is associated with more valuable patents

The above factors could serve as a set of useful patterns that can be used to evaluate whether a given patent could be considered valuable. We note that the above findings were demonstrated for patents in nanotechnology, although the considered factors are not specific to this field, but rather they concern generic information found on the cover sheet of a patent. Our future research will investigate whether contextual information that concerns nanotechnology (e.g., patent title and abstract) could be used to find additional factors that allow differentiating between successful and lower-valued patents in nanotechnology.

References

Albert, M. B., Avery, D., Narin, F., & McAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research Policy, 20*, 251–259.

Baldini, N., & Grimaldi, R. (2007). To patent or not to patent? A survey of Italian inventors on motivations, incentives, and obstacles to university patenting. *Scientometrics, 70*, 333–354.

Braun, T., Schubert, A., & Zsindely, S. (1997). Nanoscience and nanotechnology on the balance. *Scientometrics, 38*, 321–325.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Breitzman, A. F., & Mogee, M. E. (2002). The many applications of patent analysis. *Journal of Information Science, 28*, 187–205.

Carpenter, M. P., Narin, F., & Woolf, P. (1981). Citation rates to technologically important patents. *World Patent Information, 3*, 160–163.

Chen, D., Lin, W. C., & Huang, M. (2007a). Using essential patent index and essential technological strength to evaluate industrial technological innovation competitiveness. *Scientometrics, 71*, 101–116.

Chen, H., Li, X., & Lin, Y. (2007b). Worldwide nanotechnology development: a comparative study of USPTO, EPO, and JPO patents (1976–2004). *Journal of Nanoparticle Research, 9*, 977–1002.

Cohen, W. (1995). Fast effective rule induction. In *Proceedings of the twelfth international conference on machine learning* (pp. 115–123). San Mateo: Morgan Kaufmann Publishers.

Connolly, R. A., & Hirschey, M. (1988). Market value and patents: A Bayesian approach. *Economics Letters, 27*, 83–87.

Debackere, K., Verbeek, A., Luwel, M., & Zimmermann, E. (2002). Measuring progress and evolution in science and technology-II: The multiple uses of technometric indicators. *International Journal of Management Reviews, 4*, 213–231.

Domingos, P. (1999). MetaCost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth international conference on knowledge discovery and data mining* (pp. 155–164). New York: ACM Press.

Gay, C., & Le Bas, C. (2005). Uses without too many abuses of patent citations or the simple economics of patent citations as a measure of value and flows of knowledge. *Economics of Innovation and New Technology, 14*, 333–338.

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature, 28*, 1661–1707.

Gupta, V. K. (1999). Technological trends in the area of fullerenes using bibliometric analysis of patents. *Scientometrics, 44*, 17–31.

Hagedoorn, J., & Cloodt, M. (2003). Measuring innovative performance: Is there an advantage in using multiple indicators? *Research Policy, 32*, 1365–1379.

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics, 36*, 16–38.

Harhoff, D., Narin, F., Scherer, F. M., & Vopel, K. (1999). Citation frequency and the value of patented innovation. *Review of Economics and Statistics, 81*, 511–515.

Hilario, M. & Kalousis, A. (2000). Quantifying the resilience of inductive classification algorithms. In *Proceedings of the 4th European conference on principles of data mining and knowledge discovery* (pp. 106–115). France: Lyon.

Huang, Z., Chen, H., Chen, Z. K., & Roco, M. C. (2004). International nanotechnology development in 2003: Country, institution and technology field analysis based on USPTO patent database. *Journal of Nanoparticle Research, 6*, 325–354.

Huang, Z., Chen, H., Li, X., & Roco, M. C. (2006). Connecting NSF funding to patent innovation in nanotechnology (2001–2004). *Journal of Nanoparticle Research, 8*, 859–879.

Huang, Z., Chen, H., Yip, A., Ng, G., Guo, F., Chen, Z.-K., et al. (2003). Longitudinal patent analysis for nanoscale science and engineering: Country, institution and technology field. *Journal of Nanoparticle Research, 5*, 333–363.

Hullmann, A. (2007). Measuring and assessing the development of nanotechnology. *Scientometrics, 70*, 739–758.

Hullmann, A., & Meyer, M. (2003). Publications and patents in nanotechnology: An overview of previous studies and the state of the art. *Scientometrics, 58*, 507–527.

John, G. H. & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence* (pp. 338–345). San Mateo: Morgan Kaufmann Publishers.

Karki, M. (1997). Patent citation analysis: A policy analysis tool. *World Patent Information, 19*, 269–272.

Kononenko, I. (1994). Estimation attributes: analysis and extensions of RELIEF. In *Proceedings of the 1994 European conference on machine learning* (pp. 171–182). San Mateo: Morgan Kaufmann Publishers.

Kostoff, R. N., Koytcheff, R. G., & Lau, C. G. Y. (2007). Global nanotechnology research metrics. *Scientometrics, 70*, 565–601.

Kostoff, R., Stump, J., Johnson, D., Murday, J., Lau, C., & Tolles, W. (2006). The structure and infrastructure of the global nanotechnology literature. *Journal of Nanoparticle Research, 8*, 301–321.

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics, 41*, 191–201.

Lee, L. L., Chan, C. K., Ngaim, M., & Ramakrishna, S. (2006). Nanotechnology patent landscape 2006. *Nano, 1*(2), 101–113.

Leydesdorff, L., & Meyer, M. (2007). The scientometrics of a Triple Helix of university–industry–government relations. *Scientometrics, 70*, 207–222.

Lo, S.-C. (2008). Patent coupling analysis of primary organizations in genetic engineering research. *Scientometrics, 74*, 143–151.

Marinova, D., & Mcaleer, M. (2003). Nanotechnology strength indicators: International rankings based on US patents. *Nanotechnology, 14*, R1–R7.

Meyer, M. (2001). Patent citation analysis in a novel field of technology: An exploration of nano-science and nano-technology. *Scientometrics, 51*, 163–183.

Meyer, M. (2007). What do we know about innovation in nanotechnology? Some propositions about an emerging field between hype and path-dependency. *Scientometrics, 70*, 779–810.

Meyer, M., & Persson, O. (1998). Nanotechnology-interdisciplinarity, patterns of collaboration and differences in application. *Scientometrics, 42*, 195–205.

Narin, F. (1993). Technology indicators and corporate strategy. *Review of Business, 14*, 19–23.

Narin, F. (1994). Patent bibliometrics. *Scientometrics, 30*, 147–155.

Narin, F., Breitzman, A. F., & Thomas, P. (2004). Using patent citation indicators to manage a stock portfolio. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems* (pp. 553–568). Netherlands: Springer.

Narin, F., & Hamilton, K. S. (1996). Bibliometric performance measures. *Scientometrics, 36*, 293–310.

Quinlan, R. (1993). *C4.5: Programs for machine learning*. San Mateo: Morgan Kaufmann Publishers.

Reitzig, M. (2003). What determines patent value? Insights from the semiconductor industry. *Research Policy, 32*, 13–26.

Rozhkov, S., & Ivantcheva, L. (1998). Scientometrical indicators of national science & technology policy on patent statistics data. *World Patent Information, 20*, 161–166.

Sampat, B. (2004). Examining patent examination: An analysis of examiner and applicant generated prior art. Working Paper, School of Public Policy, Georgia Institute of Technology.

Tong, X., & Frame, J. D. (1992). Measuring national technological performance with patent claims data. *Research Policy, 23*, 133–141.

Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. *RAND Journal of Economics, 21*, 172–187.

Trippe, A. J. (2003). Patinformatics: Tasks to tools. *World Patent Information, 25*, 211–221.

Van Looy, B., Debackere, K., Callaert, J., Tussen, R., & Van Leeuwen, T. (2006). Scientific capabilities and technological performance of national innovation systems: An exploration of emerging industrial relevant research domains. *Scientometrics, 66*, 295–310.

Van Someren, M., & Urbancic, T. (2005). Applications of machine learning: Matching problems to tasks and methods. *Knowledge Engineering Review, 20*, 363–402.
Verbeek, A., & Debackere, K. (2006). Patent evolution in relation to public/private R&D investment and corporate profitability: Evidence from the United States. *Scientometrics, 66*, 279–294.
Verbeek, A., Debackere, K., Luwel, M., & Zimmermann, E. (2002). Measuring progress and evolution in science and technology-I: The multiple uses of bibliometric indicators. *International Journal of Management Reviews, 4*, 179–211.
Wallin, J. A. (2005). Bibliometric methods: Pitfalls and possibilities. *Basic & Clinical Pharmacology & Toxicology, 97*, 261–275.
Wang, S. (2007). Factors to evaluate a patent in addition to citations. *Scientometrics, 71*, 509–522.
Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco: Morgan Kaufman Publishers.