Enterprise Innovation with Data Mining Method Based on Naive Bayes Model Algorithm

Shan Feng

School of Computer, Hubei Polytechnic University Huangshi , China

Abstract
In the context of economic globalization, many enterprises faced considerable competition, because most business is still the traditional mode of operation, the new business innovations can be found from the business data. With the problem of huge amount of data and data redundancy, it needs a method to deal with the data processing and data mining. In this paper, the data mining based Naive Bayes model of computers knowledge is used to find new business innovations. The analysis result shows that the Naive Bayes model of data mining algorithms can effectively enhance the effective patent and new product development.
Key words: ENTERPRISE INNOVATION, NAIVE BAYES MODEL, DATA MINING

1. Introduction

Innovative concept was pioneered by Joseph Schumpeter (Schumpeter, 1912) in his book, the theory of economic development was proposed for the first time, he from the viewpoint of industrial production, that innovation is to obtain the potential for profit, implemented new combinations, namely the establishment of a new production function, production system introduced in ever new combinations of factors of production and production conditions. Chinese researchers Zhu Kigali (2008) by analyzing the similarities as well as the proposed definition against the background of innovation and that innovation concept of the broad [1], its extension can be divided into different levels. Data mining is a lot of random data which is incomplete and many noise data [2], what most people don't know, but it is also potentially useful information and knowledge in the process. Data mining is needed first collected a large amount of data in a business environment, and requires knowledge of mining is valuable. Data mining of the geometric mean calculated as follows:

$$P = \lg^{-1}\left[\frac{1}{n}(\lg x_1 + \lg x_2 + ... + \lg x_n)\right] \qquad (1)$$

2. Information requirements and analysis of innovative enterprises

Company's innovative ideas as a starting point, innovative companies are no exception. Innovative ideas are innovative companies can best reflect the stages of entrepreneurial creativity in the innovation process and innovation the key, so that with the commencement of innovation for breakthrough innovation-oriented enterprises in the idea stage of information needs is of great significance, that is from the source of innovation for a way of enterprise innovation [3], provide innovative ideas with innovation. Information needs of innovative conceptual stage include the following information.

2.1. Innovative Conception of Information Requirements

Technical information: Technical information consists of two main areas. First use of carrier technology information, symbols, graphics, video and text as a carrier of information; the second is without the carrier presents the technical information, this information is saved in the minds of staff, mainly refers to the skills, experience, and special methods. Innovation enterprise innovation conception of technical

information is enterprise technical feasibility analysis can be based on technical information, avoid duplication of development and waste, and help companies better predict the chances and risks of technological innovation [4]. In the innovation process in enterprise, the external environment is the main sources of innovative ideas, and generally includes the Internet, periodicals, exhibition and seminars, consultancy, user, upstream and downstream industry chain, scientific research institutions and universities. Technical information is particularly important for innovative enterprises, innovation is the Foundation of innovation-oriented enterprises, and depend on, if the error due to technical information innovation enterprise innovation and direction errors, then innovative companies have developed technology or products will just copy an existing product on the market. Standard deviation formula is as follows:

$$\sigma = \sqrt{\sum (x-u)^2 / N} \qquad (2)$$

Talent resources information: Can be found in the establishment and development of innovative enterprises, technology innovation-oriented enterprises have not only like traditional enterprise setting in physical capital, but is reflected in the innovation enterprise technical personnel's mind, is a creative experience and expression. Thus, innovation is the key to the success of innovation activities, is the cornerstone of enterprise innovation without innovative talent, innovation has become a fantasy. Innovative talent is the innovation of knowledge capital. Innovative enterprise's innovative talents as a top talent, scarce talent, is an epitome of enterprise's core competitiveness and innovation capability, he is very important for every innovative enterprises, innovative enterprises facing human resource is very competitive, especially in the idea stage, mistake led directly to the innovation of enterprise talent strategy innovation of enterprises in key positions don't have the talent. In the idea phase, the talent makes it easy to build its innovation team, assess the feasibility of their innovation. Standard deviation standardization is an average value minus recorded value of individual records, divided by recorded value of standard deviation:

$$X'_{ij} = \frac{x_{ij} - x_{ia}}{S_i} \qquad (3)$$

Market information: Market information is the main basis for businesses, market information including upstream and downstream industry chain information, consumer information, competitor information, and more. Markets fully reflect the situation of supply and demand, innovation-oriented enterprises to participate in the competition, innovative enterprise's marketing effectiveness, customer product reflecting the situation, market trends and other market conditions. Innovative enterprises in innovative conceptual stage, through the collection, processing, transmission, storage and use of information to determine the various kinds of marketing strategies, so that our innovative ideas in line with market rules and customer needs, survive and develop in fierce competition in the market. Thus, in production and business activities of enterprises, market information is important intangible assets of enterprises ' competition, is an important basis for enterprises to grasp market opportunities. Further, your innovative enterprise innovation is not without purpose, innovative enterprise's innovative ideas can be transformed into innovative products depends in large part on whether companies mastered the real-time dynamic of market information and market validation of the innovation enterprise innovation; he was on top of the base affect innovation enterprise innovation behavior. Let Si be the standard deviation:

$$S_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(x_{ij} - x_{ia}\right)^2} \qquad (4)$$

2.2. The Information Needs of Research and Development of Innovative

Innovation oriented enterprises to carry out development of new technologies, new methods, new technologies, new materials, new products and new market development. The information needs of research and development of innovative enterprises and innovation include:

Research and development of innovative, technical information function is to improve the efficiency of enterprise's technological innovation and at the same time reducing the uncertainty of technological development. According to the current realities of technology market development and commercial competition, innovation-oriented enterprises access to technical information and have the following two methods. One is to understand the technical information of patent documents. First, the specific patent documents reflect the details of the patent. Therefore, innovative enterprises through acquisition, reading, and analysis of patent documents, for us, our partners, competitors, the basic technology capacity of certain cognitive. Enormous innovation-oriented enterprises dependent on technical information, patent documents the most reliable information is obtained prior to introduction of the content and means. Secondly, innovation-oriented enterprises by analyzing the history of patent documents, you can learn more about the technology's life cycle stages [5]. Judge the value of imported technology for the enterprise provides an important basis. Finally, innovative enterprises

through an analysis of patent documents can know what the enterprises need patents legal status, namely technical terms of patents are protected by law. Here mainly refers to the professional scientific papers in the scientific literature. These scientific papers are typically classified and indexed. Innovative companies can use the Internet, journals and other means to obtain such professional papers. Compared to patent documents and scientific literature, the scientific literature has its own characteristics, there are three main aspects: first of all, time critical. As there is an especially great College researchers in more urgent demand for scientific literature, published in the scientific literature is timelier. The range standardization is also a common way of data standardization.

$$X_{ij} = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})} \quad (5)$$

Secondly, the accuracy of technical information more secures. Therefore, the technical reliability of the information provided in the scientific literature, scientific and accurate to a certain degree of protection. Finally, the scientific literature not only offers information related to technology innovation enterprise innovation, but also to identify the sources of relevant technologies. Enterprises through the scientific literature can be timely and accurate access to relevant sources of technical information and technology. A simplified formula of the PR value as follows:

$$PR(u) = \sum \frac{PR(V)}{L(V)} \quad (6)$$

2.3. Market Information

Enterprise of innovation is not no direction, and no purposes of innovation, but according to collected to of market information to innovation, that according to customer on products function, and appearance, and economic sex, and can operation sex, and quality, and security and effectiveness, aspects of views feedback to innovation, and on these information for height attention sex of processing, then made change, in try production stage market needs information can makes enterprise further clear itself products of market positioning, for enterprise reasonable developed products price, and improved products quality, and Improve efficiency and effectiveness of products Marketing implementation marketing information refers mainly to demand for innovation-oriented enterprises in the target market and customer information, purchase information[6], customer feedback and other information for further processing, mastering this information is conducive to innovative companies reasonably specific product marketing, promotion, and planning the next step, so as to promote product sales and sales promotion and product value.

3. Data mining and naive Bayes classification model

Humans are in the information explosion era, we were drowning in the ocean of data. in the face of a broad array of data, people tend to lose ... from a wide variety of enormous amounts of data required to obtain information is nowadays an important area of research, data mining is of great importance in this research study.

3.1. The Basic Concepts of Data Mining

In the 1960s, statisticians used terms like “Data Fishing” or “Data Dredging” to refer to what they considered the bad practice of analyzing data without an apriority hypothesis. The term “Data Mining” appeared around 1990 in the database community. For a short time in 1980s, a phrase “database mining”, was used, but since it was trademarked by HNC, a San Diego-based company, to pitch their Database Mining Workstation; researchers consequently turned to “data mining”. Other terms used include Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, etc. Gregory Shapiro coined the term “Knowledge Discovery in Databases” for the first workshop on the same topic (KDD-1989) and this term became more popular in AI and Machine Learning Community. However, the term data mining became more popular in the business and press communities. Currently, Data Mining and Knowledge Discovery are used interchangeably. Since about 2007, “Predictive Analytics” and since 2011, “Data Science” terms were also used to describe this field. Part i index collection results for group decision making is:

$$W_i = \sum_{k=1}^{m} L_k w_i^k \quad (7)$$

Data mining is a lot of, incomplete, noisy, fuzzy, in the practical application of random data, extracting hidden in it [7], people are not known in advance, but it is also potentially useful information and knowledge in the process. In this definition, required that the data source should be large, real, and noise; the discovered information and knowledge is latent and hidden behind a large amount of data, it is of interest to the user, comprehensible, and available knowledge. W index corresponds to the weight as follows:

$$W_m = (1 + \sum_{n=2}^{m} \prod_{i=n}^{m} r_i)^{-1} \quad (8)$$

So sometimes it is also known as data mining to knowledge discovery, knowledge extraction, knowledge discovery, and so on. Data mining of function and mining of target data type is related of, some function only with in a specific of data type, and some function is can application in multiple different type

of database. For data mining task of determines, must integrated considered data mining function, to mining of data type and user of interest. Data mining function for specifies data mining task in the find of mode type, its task General can is divided into two classes: Describing and forecasting. Descriptive mining tasks characterize the General properties of the data in the database. Function of data mining includes the following aspects: concept Description: characterization and differentiate between correlation analysis and classification, clustering and predictive analytics: deviation detection analysis, temporal evolution analysis and so on.

$$L_k = \frac{d_e}{\sum_{e=1}^{n} d_e} \tag{9}$$

3.2. Data Mining in Business

In business, data mining is the analysis of historical business activities, stored as static data in data warehouse databases. The goal is to reveal hidden patterns and trends. Data mining software uses advanced pattern recognition algorithms to sift through large amounts of data to assist in discovering previously unknown strategic business information. Examples of what businesses use data mining for include performing market analysis to identify new product bundles, finding the root cause of manufacturing problems, to prevent customer attrition and acquire new customers, cross-selling to existing customers, and profiling customers with more accuracy.

In today's world raw data is being collected by companies at an exploding rate. For example, Waymart processes over 20 million point-of-sale transactions every day. This information is stored in a centralized database, but would be useless without some type of data mining software to analyze it [8]. Once the results from data mining (potential prospect/customer and channel/offer) are determined, this "sophisticated application" can either automatically send an e-mail or a regular mail. Finally, in cases where many people will take an action without an offer, "uplift modeling" can be used to determine which people have the greatest increase in response if given an offer. Uplift modeling thereby enables marketers to focus mailings and offers on persuadable people, and not to send offers to people who will buy the product without an offer. Data clustering can also be used to automatically discover the segments or groups within a customer data set.

$$d_i = \sqrt{\sum_{j=1}^{m} W_j (a_{ij} - a_j)} \tag{10}$$

Businesses employing data mining may see a return on investment, but also they recognize that the number of predictive models can quickly become very large. For example, rather than using one model to predict how many customers will churn, a business may choose to build a separate model for each region and customer type. In situations where a large number of models need to be maintained, some businesses turn to more automated data mining methodologies. The distance between the elements:

$$d_i(x_k) = \sum_{i=0}^{n} (x_{ij} - M_{ij})^2 \tag{11}$$

An example of data mining related to an integrated-circuit (IC) production line is described in the paper "Mining IC Test Data to Optimize VLSI Testing." In this paper, the application of data mining and decision analysis to the problem of die-level functional testing is described. Experiments mentioned demonstrate the ability to apply a system of mining historical die-test data to create a probabilistic model of patterns of die failure. These patterns are then utilized to decide, in real time, which die to test next and when to stop testing. This system has been shown, based on experiments with historical test data, to have the potential to improve profits on mature IC products. Other examples of the application of data mining methodologies in semiconductor manufacturing environments suggest that data mining methodologies may be particularly useful when data is scarce, and the various physical and chemical parameters that affect the process exhibit highly complex interactions. Another implication is that on-line monitoring of the semiconductor manufacturing process using data mining may be highly effective.

3.3. Naive Bayes Classification Model Overview

Naive Bayes classification model is an application independent assuming that Bayes ' theorem a simple probabilistic classifier based on, it assumes that each feature is not relevant, relying on accurate natural probability models, sample concentration in supervised learning can get very good results. Naive Bayes algorithm is a method of data mining based on probability and statistics, and has a good effect in many applications. Although simple Bayesian principles and procedures are relatively simple, but its good performance can even support vector machine, SVM classifier is famous rival [9]. One of the most important areas of application is document classification. We each instance as a single document, while specific language is the subject of the document. For instance documents are news, so the class can be divided into domestic and foreign, finance, sports, entertainment and so on. Document properties are determined by key words appearing in the document's content. For example, an article "swimming" number of words a lot of documentation is the possibility of sporting

news will be greatly increased. Is a method to categorize documents with binary property indicates whether the document contains a Word. Categorize documents in this field, naive Bayesian methods have been recognized, not only because of its high processing speed and accuracy are also causes of its well-respected and the Bayes ' theorem is P:

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \quad (12)$$

3.4. Naive Bayes Principle

Naive Bayes classification model is an application independent assuming that Bayes ‘ theorem a simple probabilistic classifier based on, it assumes that each feature is not relevant, relying on accurate natural probability models, sample concentration in supervised learning can get very good results[10]. An advantage of the Naive Bayes Classifier needs only a small amount of training data.

A given dataset with many properties, calculation of p may be very large. In order to reduce the cost for p, and simplicity of the class conditional independence assumptions are made. The class label of the given sample, assuming that property value conditions of mutual independence, namely dependency property does not exist, then the:

$$P(C \mid X) = \prod_{k=1}^{n} P(x_k \mid C_i) \quad (13)$$

Probability p of which may consist of training sample valuation.

3.4. Bayesian Classification

Bell leaves republican classification is established in classic of Bell leaves republic probability theory of based of gives statistics of classification method, is a guide of learning method. Bell leaves republic classification has two classes: category is plain Bell leaves republican classification, another category is Bell leaves Republican faith network. Plain Bell leaves republic classification is a supervision of learning method. Plain Bell leaves republic classification assumes that a property of value on given class of effects independent other property value, this limit conditions strong [11], reality among cannot meet. But plain Bell leaves republic still made has is big of success, Exhibit high speed and high accuracy. With minimum classification error rate and cost overhead.

$$P(C_k \mid X) = P(X \mid C_k)P(C_k) = P(C_k)\prod_{i=1}^{n} P(X_i \mid C_k) \quad (14)$$

And Bell leaves republic network is a graphics model, can said property set between of rely on. through provides graphics of method to said knowledge, to conditions probability said variable effects of degree, through Bell leaves republic probability on a a event future may occurred of probability for estimated, overcome has based on rules of system by has of many concept and calculation of difficult, in knowledge found field has many advantages. The advantage is very good for learning! Reasoning ability can make good use of prior knowledge, the disadvantage is predicted to occur less frequently ineffective and easy learning process there is the combinatorial explosion problem.

$$\min = \sum (p_i - (ap_i + b))^2 \quad (15)$$

4. Applying Bayesian model in data mining of enterprise innovation point

In a Naive Bayes model of the data mining algorithms help when looking for innovation, we have statistics on innovation promotion in various industries. Figure 1 shows the sample enterprises are engaged in the distribution of the type of industry or industry descriptions. You can see in figure, production-oriented enterprises a proportion of 55.64%, trading for 9.89%, information technology for 14.79% services for 19.68%, we can see that enterprise industry is dominated by manufacturing and service industries.

After applying our algorithm, can see the increase of innovations can significantly improve the average income. It shows our Bayesian data mining algorithms can achieve very good results in practical application, is beneficial to the economic development of enterprise development. Figure 2 displayed sample enterprise near two years average sales income in industry in the of relative location [12], which is located in industry partial level of only accounted for 13.31%, is located in industry average, and below industry average and industry average has larger gap of total 86.69%, showed that was survey enterprise are in venture stage, enterprise scale in medium and lower level, to SMEs mostly, this and said to number for scale survey of situation basic match.

At present, up to national strategy will innovation perspective, States the attached great importance to independent innovation. First talent for innovation, advancement of the globalized economy of the 21st century is the era of the knowledge economy, knowledge of competition is competition for talent, if an enterprise has what one needs professionals, has

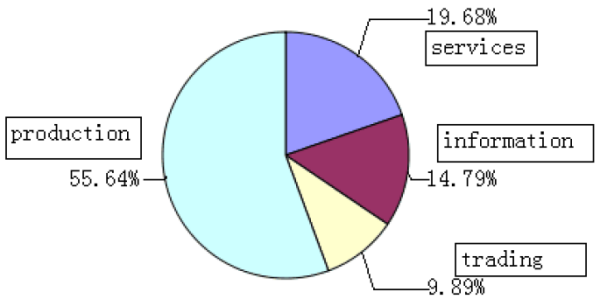

Figure 1. Promote Industry Innovation

been winning in the competition. Secondly, characterized by the increase in the number of innovative enterprises, is the subject of innovation activities of enterprises, innovative enterprises directly determine the number of innovative activity of the whole society. Since entering the new century, our country has R&D number of projects increased, 7,116 growth from 2000 to 2010 12,889, growth rates remained at a relatively stable state.

With the continuous economic and social development, national attention increasingly on innovation, innovations are significant. But because of the character of innovation, not many indicators to measure their results, new product development and the effective number of invention patents are the two most direct measures. And our Naive Bayes model of data mining algorithms can effectively enhance the effective patent and new product development.

5. Conclusions

In recent years, as China's economy with the world, enterprises are increasingly faced the risk of being eliminated by the competition. Enterprises want long-term success, it must have its own core competence, and this must be based on the core competence and innovation as premise. Corporate governance is the Foundation of company system, naturally would have a negligible impact on innovation. Based on the review on the basis of previous research, seek out vulnerabilities and entry point for the present study, by updating the definition of corporate governance, the introduction of external governance structure, and build a new framework in the study on Enterprise Innovation behavior based on corporate governance. In the information age we need things applied to enterprise innovation in the field of the computer industry, and enhance the competitiveness of enterprises to enable them to adapt to the development, and enhance their core competitiveness rely on innovations. Our Naive Bayes algorithm can meet the company's needs, to help them upgrade their ability to innovate.

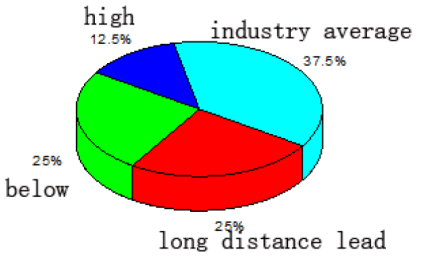

Figure 2. Promote Benefits of Industry Innovation

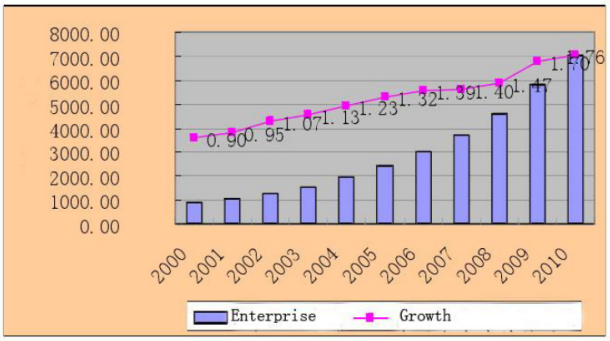

Figure 3. Innovative Enterprise Growth Curve

References

1. Datcu M.,Seidel K. "Image information mining: exploration of image content in large archives"[C].IEEE Aerospace Conference Proceedings,2000(3):253~264
2. Klose Aljoscha, Kruse Rudolf, Gross Hermann. Tuning on the fly of structural image analysis algorithms using data mining[C].Proceedings of SPIE-The International Society for Optical Engineering,2000.4055:311~321.
3. Zequn, G. Scan patterns for association rule mining of image data[C].Proceedings of SPIE-The International Society for Optical Engineering,2003(4898):212~219.
4. Dietterich TG. Machine learning research: four current directions[J]. AI Magazine, 1997,18(4): 97-136
5. Zhou Z H, Wu J, Tang W. "Ensembling neural networks: many could be better than all"[J]. Artificial intelligence, 2002, 137(1): 239-263.
6. ChengJ,Greiner R. Comparing Bayesian networke classifiers. In: Laskey KB, PradeH, eds. Proe. Of the 15th Conf.on Uneertainty in Artifieial Intelligenee. SanFraneiseo: Morgan Kaumfann Publishers,1999.101-108.
7. David Heeherman. "A Tutorial on Learning with Bayesian Networks" [R]. Mierosoft Researeh,1995.
8. Kaizhu Huang, Irwin King, Miehael R.Lyu. Learning Maximum Likelihood Semi-Naïve Bayesian Network Classifier.2002
9. Gautam Ahuja, Riitta Katila. Technological Acquisition and The Innovative Performance of Acquiring Firms: A Longitudinal Study[J]. Strategic Management Journal, 2001,22(3): 197-220.
10. PAK tee NG. "The learning organization and the innovative organization "[J]. HumanSystems Management, 2004(23): 93-100.
11. CAMELO-ORDAZ C,FERNANDEZ-ALLES M D, LALUZ,et al. Top management team tision and human resources management practices in innovative Spanish companies[J] . International Journal of Human Resource Management, 2008,19(4): 620-638.
12. Hall.The Financing of Research and Development[J].Oxford Review of Economic Policy, 2002(1):35- 51.