Data Mining Technique for Knowledge Discovery from Engineering Materials Data Sets

Doreswamy[1], K.S. Hemanth [2], Channabasayya M. Vastrad[3], and S. Nagaraju[4]

[1,2] Department of Computer Science
Mangalore University, Mangalagangotri-574 199, Karnataka, India
doreswamyh@yahoo.com, reachhemanthmca@gmail.com
[3] Department Computer Science & Engineering
PDIT, Hospet, Karnataka, India
chennu.vastrad@gmail.com
[4] Department Computer Science & Engineering
Bahubali College of Engineering, Shravanabelagola,
Hassan-573135, Karnataka, India
nagaraju.sms@gmail.com

Abstract. The goal of this paper is to discuss how data mining technique can be applied in materials informatics to extract knowledge from materials data. Studying material data sets from a data mining perspective can be beneficial for manufacturing and other industrial engineering applications. This work employs an effective materials classification system on design requirements. Experiments were conducted on material datasets that consist of all class of materials. The algorithm of the Naive Bayesian classifier is implemented successively enabling it to solve classification problems and the outcomes can be very useful for design engineers to speed up decision making process in manufacturing and other industrial engineering applications. The comparison of performance with various domains of material classes confirms the advantages of successive learning and suggests its application to other learning domains.

Keywords: Knowledge Discovery, Materials informatics, Naive Bayesian Classifier.

1 Introduction

The rapid developments in materials science and information technologies have influenced the large volume of massive data sets and materials informatics respectively. Materials informatics a field of study that applies the principles of informatics to materials science and engineering to better understand the use, selection, development, and discovery of materials.

As a lot of traditional analytic techniques employed for materials structural-properties analysis and not effective any longer under these situations, researchers in the manufacturing industries and other industrial engineering applications areas are being faced more new research issues in systematic analysis of materials data sets. Therefore, materials informatics has been emerging in material science and technology as a new

N. Meghanathan et al. (Eds.): CCSIT 2011, Part I, CCIS 131, pp. 512–522, 2011.

research areas[5],[10],[11], and has already changed the experimental methods and way of thinking in materials research, and will lead even more challenges in interdisciplinary research.

Data Mining is an interdisciplinary field merging ideas from statistics, machine learning, information science, visualization and other disciplines[7]. It is a very useful approach to integrate information and theory for knowledge discovery from any informatics such Bioinformatics, Chemoinformatics, Nano informatics, Materials informatics and so on. The impact of Data Mining and knowledge discovery has been evidenced by many successful research experimental results[19],[20],[21],[22]. Therefore, Data mining can be used to extract non-trivial, hidden, potential useful and ultimately understandable knowledge from massive materials databases[29],[30].

Data Mining has two primary Models: Descriptive Data Mining Model and predictive Data Mining Model. Descriptive mining models describe or summarize the general characteristics or behaviour of the data in the materials database. Predictive models perform inference on the current data in order to make the prediction. Both of them are fundamentals to understand materials behaviors. In general , in materials informatics, Data mining can be used in the following task[13]:

(i). **Association analysis:** Association analysis is good at discovering patterns, and can be used to develop heuristic rules for materials behaviour based on large data sets[26],[27],[28].

(ii). **Classifier/Predict modelling:** Some machine learning algorithms can be used for materials class prediction and materials classification models such as support vector regression (SVR) and neural network (NN), can be used to build up the Predict models[31]. These models can be used to predict crystal structure or composite materials properties from fused materials data[14].

(iii). **Cluster analysis:** As an exploratory data analysis tool, it can sort different materials or properties into groups in such a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. And, cluster analysis can be integrated with high-throughput experimentation for rapidly screening combinatorial data[20].

(iv). Outlier Analysis: In properties analysis or combinatorial experiments, outlier analysis is used to identify anomalies, especially to assess the uncertainty and accuracy of results, and distinguish between true discoveries and false-positive results.

(v). **Material visualization:** Reconstruction of material structure information based on materials data would help researchers to analyze the relationships between material structure and material properties[16].

The rest of the paper is organized as follows: scope of knowledge discovery on materials informatics is discussed in section 2. The section 3 describes naive Bayesian classifier algorithm and performance measures. The experimental results are presented in the section 4. The conclusions and future scopes are given in the section 5.

2 Scope of Data Mining in Materials Informatics

Data Mining is becoming an increasingly valuable tool in the broad area of materials development and design[4],[9], and there are good reasons why this area is particularly rich for materials informatics[14],[15],[16],[17]. There is a massive range of possible new materials, and it is often complex to physically model the relationships between constituents, processing and final properties. Therefore, materials are primarily still developed by quantitative and trial-and-error procedures, where researchers are guided by experience and heuristic rules for materials classification, selection and property predictions. These rules are applied to somewhat limited materials data sets of constituents and processing conditions, but then try as many combinations as possible to find materials with desired properties. This is essentially human Data Mining, where one's brain, rather than the computer, is being used to find correlations, make predictions, and design optimal strategies. Transferring Data Mining tasks from human to computer offers the potential to enhance accuracy, handle more data, and to allow wider dissemination of accrued knowledge[5].

Materials informatics[18],[19] has been a subject of materials science, since the international conference of "Materials Informatics-Effective Data Management for New Materials Discovery" was held in Boston in 1999. Wei[24] described that materials informatics is a new subject that leverages information technology and computer network technology to represent, parse, store, manage and analyze the material data, in order to realize the sharing and knowledge mining of materials data for uncovering the essence of materials, and accelerate the new material discovery and design. The research areas of materials informatics are mainly focused on following tasks[25]:

i. **Data standards:** There are thousands of materials databases in different formats[32], and they are difficult to communicate with each other. To standard these databases and to integrate materials data into a single or coherent database, data pre-processing is the first important task of materials informatics[23] to enable knowledge discovery.

ii. **Organization and management of material data:** In order to meet materials researchers' different needs, satisfy the need of research and production, to construct the materials data into a whole and single coherent database, efficient Materials Database Management Systems(MDBMS)is very necessary[25].

iii. **Data mining on materials data:** There is an enormous range of possible new materials, and it is often difficult to physically model the relationships between constituents, and processing, and final properties. Data mining has the abilities to search, classify, select and analyze material data and to find potential, previously unknown patterns rules. It involves selecting, exploring and modelling large amounts of data to uncover previously unknown patterns from large materials databases[7]. Data mining involves some high-effective computational algorithms[18],[19], such as neural networks, genetic algorithm, etc.

3 Data Mining Technique

Classification and prediction is one of the core tasks of Data Mining. A classification technique is a systematic approach to building classification models from input data set. Several classification models are reported in literature such include Decision Tree Classifier, Rule-Based Classifier, Neural Network Classifier, naive Bayesian Classifier, Neuro-fuzzy classifier, Support Vector Machines and etc. Each technique employs a learning algorithm to identify a model that best fits the relationships between the attribute set and class label of the input data. The model generated by a learning algorithm should both fit the input data well and correctly predict the class labels of data set that has never seen before. Therefore, the key objective of the learning algorithm is to build models with good generalization capacity.

3.1 Naive Bayesian Classifier

Naive Bayesian classifier is a statistical classifier that can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. It is fast and incremental that can deal with discrete and continuous attributes and has excellent performance in real-life problems. The naive Bayesian classifier, or simple Bayesian classifier generally used for classification or prediction task. As it is simple, robust and generality, this procedure has been deployed for various applications such as Materials damage detection[1],[2], Agricultural land soils classification[6], Network intrusion detection[8], Machine learning applications[19]. Therefore, the application of this method is extended to classification of engineering materials data sets[4],[6],[9] and to reduce the computational cost of classification of materials. The working procedure of the naive Bayesian classifier is shown in the followings section.

3.2 Algorithm of Naïve Bayesian Classifier

1. Let D be a training set of tuples and their associated class labels. Each tuple is represented by an n-dimensional attribute vector, $X = (x_1, x_2, ...x_n)$, depicting n measurements made on the tuple from n attributes, respectively, $A_1, A_2, ...A_n$.
2. Suppose that there are *m* classes, $C_1, C_2, ...C_m$. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class C_i, if and only if

$$\mathrm{P}\ (C_{\mathrm{i}}/\mathrm{X}) > \mathrm{P}\ (C_j/\mathrm{X}) \text{ for all } 1 \le \mathrm{j} \le \mathrm{m}; \mathrm{j} \ne \mathrm{i}. \quad (1)$$

3. Thus it maximizes $\mathrm{P}\ (C_{\mathrm{i}}/\mathrm{X})$. The class C_i for which $\mathrm{P}\ (C_{\mathrm{i}}/\mathrm{X})$ is maximized is called the maximum posteriori hypothesis. By Bayes' theorem .

$$\mathrm{P}\ (C_{\mathrm{i}}/\mathrm{X}) = \frac{P(X / C_i)P(C_i)}{P(X)} \quad (2)$$

As $P(X)$ is constant for all classes, only $P(X / C_i)P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_i) = P(C_2) = P(C_3) = ... = P(C_m)$, and it would therefore maximize $P(X/C_i)$. Otherwise, it maximizes $P(X/C_i)P(C_i)$. Note that the class prior probabilities may be estimated by $P(C_i) = |C_i, D| / |D|$, where $|C_i, D|$ is the number of training tuples of class C_i in D.

4. Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X/C_i)$. In order to reduce computation in evaluating $P(X/C_i)$, the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

$$P(X/C) = \prod_{k=1}^{n} P(X_k/C_i)$$

$$= P(X_1/C_i) \times P(X_2/C_i) \times P(X_2/C_i) \times ....P(X_n/C_i) \qquad (3)$$

The probabilities $P(X_1/C_i), P(X_2/C_i), P(X_3/C_i)$...... $P(X_n/C_i)$ can easily be estimated from the training tuples. Recall that here x_k, refers to the value of attribute A_k, for tuple X. For each attribute, the attribute value may be either categorical or continuous-valued. For instance, to compute $P(X/C_i)$, it is considered the following:

If A_k is categorical, then $P(X_k/C_i)$ is the number of tuples of class C_i in D having the value for A_k, divided by $|C_i, D|$, the number of tuples of class C_i in D.

5. In order to predict the class label of X, $P(X/C_i)P(C_i)$ is evaluated for each class C_i. The classifier predicts that the class label of tuple X is the class C_i if and only if

$$\mathrm{P}\ (\mathrm{X/C}_{\mathrm{i}})P(C_i) > \mathrm{P}\ (\mathrm{X/C}_{\mathrm{j}})\ \ \mathrm{P}\ (\mathrm{C}_{j})\ \text{ for all } 1 \le \mathrm{j} \le \mathrm{m}; \mathrm{j} \ne \mathrm{i} \qquad (4)$$

In other words, the predicted class label is the class C_i for which $P(X/C_i)P(C_i)$ is the maximum.

3.3 Performance Measures

The classifier in this research is evaluated on engineering materials data set using the standard metrics of accuracy, precision, and recall. These were calculated using the predictive classification table, known as Confusion Matrix[16].

Table 1. Confusion Matrix

		PREDICTED	
		IRRELEVANT	RELEVANT
ACTUAL	IRRELEVANT	TN	FP
	RELEVANT	FN	TP

Where:

TN (True Negative) : Number of correct predictions that an instance is irrelevant

FP (False Positive) : Number of incorrect predictions that an instance is relevant

FN (False Negative) : Number of incorrect predictions that an instance is irrelevant

TP (True Positive) : Number of correct predictions that an instance is relevant

Accuracy – The proportion of the total number of predictions that were correct:

$$\text{Accuracy } (\%) = (TN + TP) \,/\, (TN + FN + FP + TP) \qquad (5)$$

Precision – The proportion of the predicted relevant materials data sets that were correct:

$$\text{Precision } (\%) = TP \,/\, (FP + TP) \qquad (6)$$

Recall – The proportion of the relevant materials data sets that were correctly identified

$$\text{Recall } (\%) = TP \,/\, (FN + TP) \qquad (7)$$

The classification performance of the naive Bayesian classifier is analyzed with standard metrics in the experimental results.

4 Experimental Results and Discussion

In this experiment, materials database is organized by sampling material data sets from peer-reviewed research papers published[23] and from poplar materials website http://www.matweb.com. The tuples of data table consists of both numerical and categorical attribute values. The tuples consisting only categorical attributes and their values are considered for finding probable class of materials. Atypical set of training sample data sets is shown in the following table 2.

A prototype software module realizing Naive Bayesian classifier is designed and developed using .NET technology as it is efficient for handling data objects. This

Table 2. Partial List Of Training Samples And Their Attribute Values

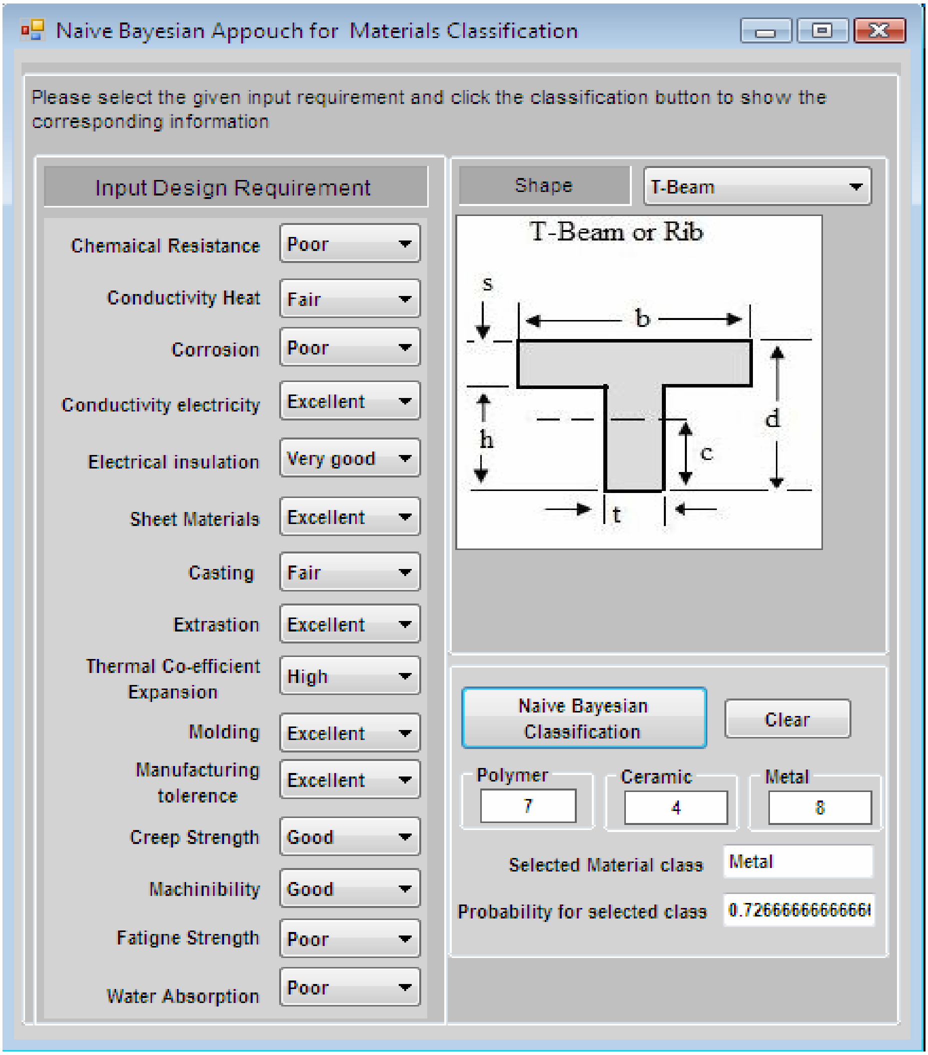
CR	CH	CE	SM	CAST	EXTRN	MANFT	CS	MACHI	FS	WA	Class Label
Excellent	*Poor*	*NIL*	*Good*	*Fair*	*Good*	*Excellent*	*Poor*	*Good*	*Poor*	*Poor*	*P*
Good	*Poor*	*NIL*	*Good*	*Fair*	*Good*	*Excellent*	*Poor*	*Good*	*Poor*	*Poor*	*P*
Good	*Poor*	*NIL*	*Good*	*Fair*	*Good*	*Excellent*	*Poor*	*Good*	*Poor*	*Poor*	*P*
Good	*Poor*	*NIL*	*Good*	*Fair*	*Good*	*Excellent*	*Poor*	*Good*	*Poor*	*Poor*	*P*
Very Good	*Poor*	*NIL*	*Good*	*Fair*	*Good*	*Excellent*	*Poor*	*Good*	*Poor*	*Poor*	*P*
Excellent	*Poor*	*Good*	*Poor*	*Poor*	*Poor*	*Good*	*Excellent*	*Poor*	*Good*	*Poor*	*C*
Excellent	*Poor*	*Good*	*Poor*	*Poor*	*Poor*	*Good*	*Excellent*	*Poor*	*Good*	*Poor*	*C*
Good	*Poor*	*Good*	*Poor*	*Poor*	*Poor*	*Good*	*Excellent*	*Poor*	*Good*	*Poor*	*C*
Good	*Fair*	*Good*	*Poor*	*Poor*	*Poor*	*Good*	*Excellent*	*Poor*	*Good*	*Poor*	*C*
Good	*Fair*	*Good*	*Poor*	*Poor*	*Poor*	*Good*	*Excellent*	*Poor*	*Good*	*Poor*	*C*
Poor	*Very Good*	*Excellent*	*Excellent*	*Excellent*	*Excellent*	*Fair*	*Good*	*Good*	*Good*	*Poor*	*M*
Poor	*Good*	*Excellent*	*Excellent*	*Excellent*	*Excellent*	*Fair*	*Good*	*Good*	*Good*	*Poor*	*M*
Good	*Good*	*Excellent*	*Excellent*	*Excellent*	*Excellent*	*Fair*	*Good*	*Good*	*Good*	*Poor*	*M*
Fair	*Good*	*Excellent*	*Excellent*	*Excellent*	*Excellent*	*Fair*	*Good*	*Good*	*Good*	*Poor*	*M*
Poor	*Good*	*Excellent*	*Excellent*	*Excellent*	*Poor*	*Fair*	*Good*	*Good*	*Excellent*	*Fair*	*M*
Good	*Fair*	*Good*	*Poor*	*Poor*	*Fair*	*Good*	*Excellent*	*Poor*	*Good*	*Fair*	*C*
Good	*Fair*	*Good*	*Poor*	*Poor*	*Fair*	*Good*	*Excellent*	*Poor*	*Good*	*Fair*	*C*
Poor	*Very Good*	*Good*	*Excellent*	*Excellent*	*Very Good*	*Fair*	*Good*	*Good*	*Good*	*Good*	*M*
Poor	*Good*	*Good*	*Excellent*	*Excellent*	*Good*	*Fair*	*Good*	*Good*	*Poor*	*Good*	*M*
Good	*Poor*	*Good*	*Good*	*Fair*	*Good*	*Excellent*	*Poor*	*Good*	*Poor*	*Fair*	*P*

CR: Chemical Resistance, CH: Conductivity-Heat, CE: Conductivity-Electricity SM: Sheet Metal, CAST: Casting, EXTRN: Extrusion, MANFT: Manufacturing, CS: Creep Strength, MACHN: Mach inability, FS: Fatigue Strength, WA: Water Absorptions

software module accepts design requirements from the user's Graphical User Interface(GUI) and predicts probable class to which design requirements belong. The design requirements associated geometrical features of the material are determined by design engineers or through CAD systems. The GUI of the implemented software module is shown in the following figure 1.

The Naive Bayesian classifier is trained on engineering materials data set consisting of 1630 data sets and these data sets consist of 15 discrete and categorical attributes. A sample data set is randomly selected from the testing data set and input to the classier, then classifier predicts the probable knowledge, class of the input data set. Misclassification occurs when an input data sample's categorical attribute values neither associated to any of the class of materials. The classification performance of naive Bayesian classifier is analyzed with the standard measures, which are used for measuring the other classifiers, shown in the figures 2 and 3. From the figure 2, the False Negative(FN) and False Positive(FP) measures in all the class of materials are very less that indicate the false data sets correctly classified by the classifier.

The knowledge extracted by the classifier includes the general features of the tested data sets, total number of data sets undergone for testing, number of data sets classified to each class, number of true data sets positively classified, number of true data sets negatively classified, number of false data sets positively classified and number of false data sets negatively classified, are visualized in the figure 2.



Naive Bayesian Appouch for Materials Classification

Please select the given input requirement and click the classification button to show the corresponding information

Input Design Requirement

Chemaical Resistance	Poor
Conductivity Heat	Fair
Corrosion	Poor
Conductivity electricity	Excellent
Electrical insulation	Very good
Sheet Materials	Excellent
Casting	Fair
Extrastion	Excellent
Thermal Co-efficient Expansion	High
Molding	Excellent
Manufacturing tolerence	Excellent
Creep Strength	Good
Machinibility	Good
Fatigne Strength	Poor
Water Absorption	Poor

Shape: T-Beam

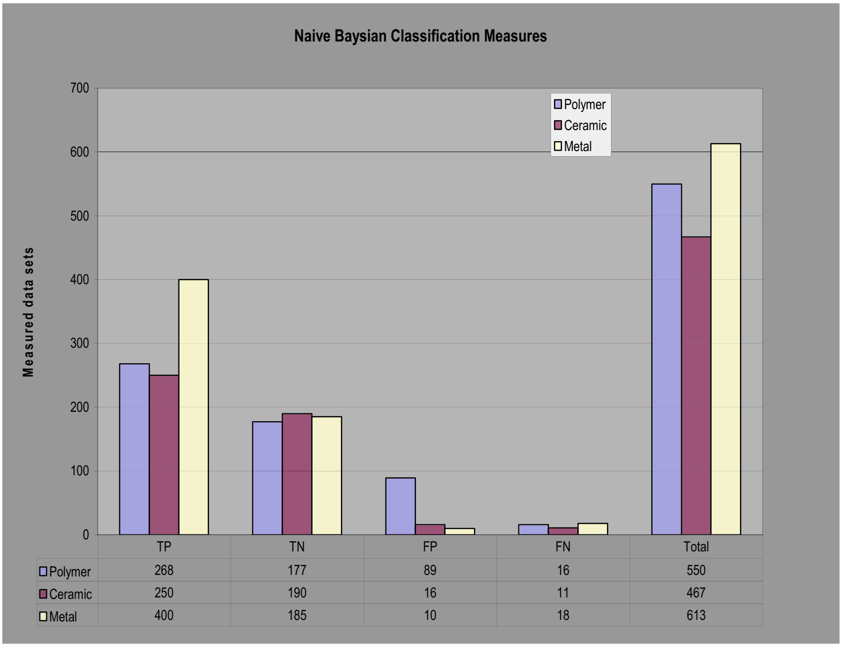
Polymer	Ceramic	Metal
7	4	8

Selected Material class: Metal

Probability for selected class: 0.7266666666666

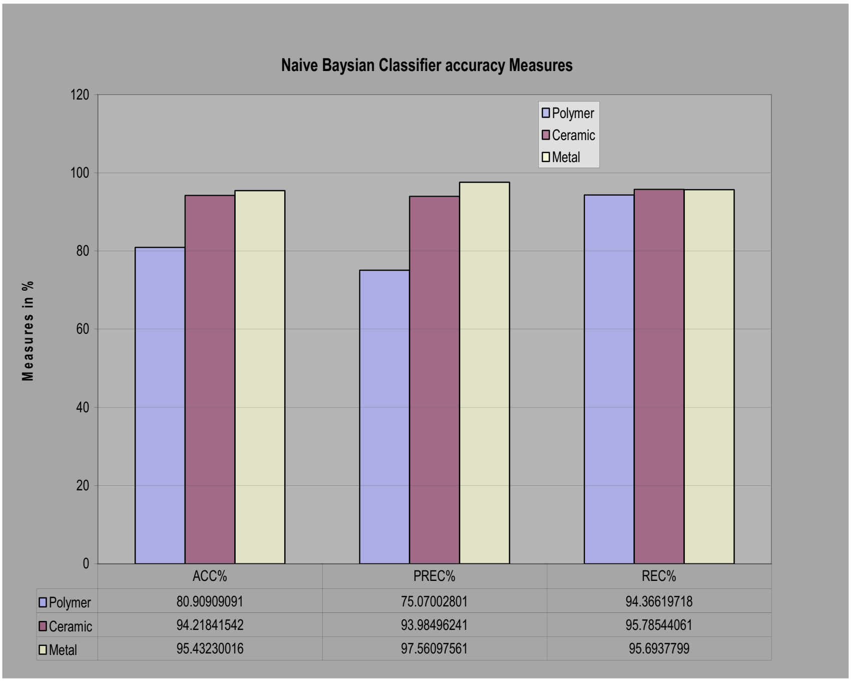
Fig. 1. A prototype software module for probable material's class prediction on input design requirements

The aggregated features computed with obtained general features include the performance measures such as Accuracy(ACC), Precision(PREC) and Recall (REC) and these are depicted in figure 3. ACC, PREC and REC are computed on 1630 data sets. Out these, ACC is 80.91% , PREC is 75.07% and REC is 94.37 % for 550 polymer materials. **ACC is** 94.22%, PREC is 93.98% and REC is 95.78% for 467 Ceramic materials, and **ACC is** 95.43,%, PREC is 97.56% and REC is 95.69% for 617 metals materials.



	TP	TN	FP	FN	Total
□ Polymer	268	177	89	16	550
■ Ceramic	250	190	16	11	467
□ Metal	400	185	10	18	613

Fig. 2. Classification measures of TP,TN,FP,FN and total number of materials in each class



	ACC%	PREC%	REC%
□ Polymer	80.90909091	75.07002801	94.36619718
■ Ceramic	94.21841542	93.98496241	95.78544061
□ Metal	95.43230016	97.56097561	95.6937799

Fig. 3. Classification performance measures ACC, PREC, and REC

5 Conclusions and Future Scope

In this paper, Data Mining technique is applied in materials informatics to extract knowledge from materials data. The performance of the naive Bayesian classifier is analyzed on materials data sets. Studying material data sets from a data mining perspective can be beneficial for manufacturing and other industrial engineering applications. The algorithm of the Naive Bayesian classifier is applied successively enabling it to solve classification problems and the outcomes can be very useful for the manufacturing and other industrial engineering applications. The comparison of performance in various domains of material classes confirms the advantages of successive learning and suggests its application to other learning algorithms.

Further, an application of this algorithm can be extended to the classification of engineering materials data sets consisting of both numerical and categorical attribute values. Performance comparison of this algorithm with other classification algorithms on materials informatics data sets is the future scope of this research.

Acknowledgement. The authors wish to acknowledge the financial support from the University Grant Commission(UGC), INDIA for the Major Research Project **"Scientific Knowledge Discovery Systems (SKDS) For Advanced Engineering Materials Design Applications"** vide reference F.No. 34-99\2008 (SR), 30th December 2008, and also gratefully acknowledge the unanimous reviewers for their kind suggestions and comments for improving this paper.

References

[1] Addin, O., Sapuan, S.M., Mahdi, E., Othman, M.: A Naive-Bayes classifier for damage detection in engineering, materials. Materials and Design, 2379–2386 (2007)

[2] Addina, A.O., Sapuanb, S.M., Othmanc, M.: A Naïve-Bayes Classifier And F-Folds Feature Extraction Method For Materials Damage Detection. International Journal of Mechanical and Materials Engineering (IJMME) 2(1), 55–62 (2007)

[3] Bhargavia, P., Jyothi, S.: Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils. International Journal of Computer Science and Network Security 9(8), 117–122 (2009)

[4] Doreswamy, Sharma, S.C.: An Expert Decision Support System for Engineering Materials Selections And Their Performance Classifications on Design Parameters. International Journal of Computing and Applications (ICJA) 1(1), 17–34 (2006)

[5] Doreswamy: A survey for Data Mining framework for polymer Matrix Composite Engineering materials Design Applications. International Journal of Computational Intelligence Systems (IJCIS) 1(4), 312–328 (2008)

[6] Doreswamy: Engineering Materials Classification Model- A Neural Network Application. IJDCDIS A Supplement, Advances in Neural Networks 14(S1), 591–595 (2007)

[7] Han, J., Kamber, M.: Data Mining Concepts and Techniques. Morgan Kaufmann Publisher, San Francisco (2009)

[8] Khor, K.-C., Ting, C.-Y., Amnuaisuk, S.-P.: From Feature Selection to Building of Bayesian Classifiers: A Network Intrusion Detection Perspective. American Journal of Applied Sciences 6(11), 1949–1960 (2009)

[9] Langseth, H., Nielsen, T.: Classification using Hierarchical Naïve Bayes models. Machine Learning 63(2), 135–159 (2006)

[10] Chikyow, T.: Trends In Materials Informatics In Research On Inorganic Materials. Quarterly Review 20, 59–71 (2006)

[11] Qunyi, W., Xiaodong, P., Xiangguo, L., Weidong, X.: Materials informatics and study on its further development. Chinese Science Bulletin 51(4), 498–504 (2006)

[12] Callister, W.D.: Materials Science and Engineering. Wiley India Pvt. (2007)

[13] Suh, C., Rajan, K.: Data mining and informatics for crystal chemistry: establishing measurement techniques for mapping structure-property relationships. Materials Science And Technology 25, 466–471 (2009)

[14] Scott, D. J., Coveney, P.V., Kilner, J. A., Rossiny, J., Alford, N.: Prediction of the functional properties of ceramic materials from composition using artificial neural networks. Journal Of The European Ceramic Society 27, 4425–4435 (2007)

[15] Ferris, F., Peurrung, L.M., Marder, J.: Materials informatics: Fast track to new materials. Advanced Materials & Processes 165, 50–51 (2007)

[16] Yu., G., Chen., J., Zhu, L.: Data mining techniques for materials informatics: Datasets Preparing and Applications. In: Proc. 2009 Second International Symposium on Knowledge Acquisition and Modeling, pp. 181–189 (2009)

[17] Rodgers, J.R., Cebon, D.: Materials informatics. MRS Bulletin 31, 975–980 (2006)

[18] Rodgers, J.R.: Materials informatics: Knowledge acquisition for materials design. Abstracts Of Papers Of The American Chemical Society 226, 302–303 (2003)

[19] Rajan, K.: Combinatorial materials sciences: Experimental strategies for accelerated knowledge discovery. Annual Review Of Materials Research 38, 299–322 (2008)

[20] Rajan, K.: Combinatorial materials sciences: Experimental strategies for accelerated knowledge discovery. Annual Review Of Materials Research 38, 299–322 (2008)

[21] Moliner, M., Serra, J.M., Corma, A., Argente, E., Valero, S., Botti, V.: Application of artificial neural networks to high-throughput synthesis of zeolites. Microporous and Mesoporous Materials 78, 73–81 (2005)

[22] Fischer, C., Tibbetts, K.J., Morgan, D., Ceder, G.: Predicting crystal structure by merging data mining with quantum mechanics. Nature Materials 5, 641–646 (2006)

[23] Song, Q.G.: A preliminary investigation on materials informatics. Chinese Science Bulletin 49, 210–214 (2004)

[24] Wei, Q.Y., Peng, X.D., Liu, X.G., Xie, W.D.: Materials informatics and study on its further development. Chinese Science Bulletin 51, 498–504 (2006)

[25] Hrubiak, R., George, L., Saxena, S.K., Rajan, K.: A Materials Database for Exploring Material Properties. Journal of Materials 61, 59–62 (2009)

[26] Rajan, K.: Informatics and Integrated Computational Materials Engineering: Part II. JOM 61, 47–47 (2009)

[27] Broderick, S., Suh, C., Nowers, J., Vogel, B., Mallapragada, C., Narasimhan, B., Rajan, K.: Informatics for combinatorial materials science. JOM 60, 56–59 (2008)

[28] Takeuchi, I., Lippmaa, M., Matsumoto, Y.: Combinatorial experimentation and materials informatics. MRS Bulletin 31, 999–1003 (2006)

[29] Hunt, W.H.: Materials informatics: Growing from the bio world. JOM 58, 88–88 (2006)

[30] Inmon, W.H.: Building the Data Warehouse, 4th edn. John Wiley and Sons, Inc., New York (2007)

[31] Li, Y.: Predicting materials properties and behaviour using classification and regression trees. Materials Science And Engineering A-Structural Materials Properties Microstructure And Processing 433, 261–268 (2006)

[32] Westbrook, J.H.: Materials Data On The Internet. Data Science Journal 2(25), 198–211 (2003)