Mining Knowledge of the Patient Record: "The Bayesian Classification to Predict and Detect Anomalies in Breast Cancer"

Souad Demigha
CRI, Sorbonne University, Paris, France
souad_demigha@yahoo.fr

Abstract: knowledge management, data mining, and text mining techniques have been adopted in various successful biomedical applications in recent years. Data Mining (DM) is the most important subfields in knowledge management (KM). It has been proven that data mining can enhance the KM process with better knowledge. In this paper, we investigate the application of DM techniques for mining knowledge of the patient record. The patient record represents documents of the patient's examinations and treatments. Data Mining is the process of "mining" or extracting information from a data set and transform it into an understandable structure for further use. We propose a methodology for mining medical knowledge based on the Bayesian Classification to predict and detect anomalies in breast cancer. We use the Naïve Bayes Algorithm to develop this methodology. We illustrate the knowledge mining process by real examples of medical field. We investigate through these illustrations how knowledge is better mined and thus, reused when applying concepts and techniques of Data Mining. On the other hand, we investigate the potential contribution of the Naive Bayesian Classification methodology as a reliable support in computer-aided diagnosis of such events, using the well-known Wisconsin Prognostic Breast Cancer dataset. Finally, we will demonstrate the suitability and ability of the Naive Bayes methodology in Classification/Prediction problems in breast cancer.

Keywords: Patient Record, Data Mining, Bayesian Classification, Naïve Bayes Algorithm, Breast cancer prediction

1. Introduction

Knowledge management (KM), data mining (DM), and text mining techniques have been adopted in various successful biomedical applications in recent years. Knowledge management techniques and methodologies have been used to support the storing, retrieving, sharing, and management of multimedia and mission-critical tacit and explicit biomedical knowledge. Data mining techniques have been used to discover various biological, drug discovery, and patient care knowledge and patterns using selected statistical analyses, machine learning, and neural networks methods. Text mining techniques have been used to analyse research publications as well as electronic patient records (Chen, Fuller, Friedman and Hersh, 2005).

There is no universal definition of KM. KM could be defined as the process of collecting and identifying useful information (knowledge acquisition), transferring tacit knowledge to explicit knowledge (knowledge transfer or creation), sharing the knowledge in the repository (organizational memory), disseminating it through the whole organization (knowledge sharing), enabling employees to easily retrieve it (knowledge retrieval) and exploiting and usefully applying knowledge (knowledge leverage), (Demigha$_a$, 2015). In the biomedical context, knowledge management practices need to influence existing clinical decision support, information retrieval, and digital library techniques to capture and deliver tacit and explicit biomedical knowledge, (Chen, Fuller, Friedman and Hersh, 2005).

Data Mining refers to "extracting" or "mining" knowledge from large amounts of data (Han and Kambar, 2006). Data mining is an essential part of knowledge management (KM). It has been proven that data mining can enhance the KM process with better knowledge. Wang and Wang (2008) point that data mining can be useful for KM to share common knowledge of business intelligence context among data miners and to use data mining as a tool to extend human knowledge. Data mining tools could help organizations to discover the hidden knowledge in the enormous amount of data, (Wang and Wang, 2008).

Data mining is often used during the knowledge discovery process and is one of the most important subfields in knowledge management. Data mining aims to analyse a set of given data or information in order to identify novel and potentially useful patterns (Fayyad, Piatetsky-Shapiro and Smyth, 1996).These techniques, such as Bayesian models, decision trees, artificial neural networks, associate rule mining, and genetic algorithms, are often used to discover patterns or knowledge that are previously unknown to the system and the users (Dunham, 2002), (Chen and Chau, 2004). Data mining has been used in many applications such as marketing, customer relationship management, engineering, medicine, crime analysis, expert prediction, Web mining, and mobile computing, among others.

Reference this paper as: Demigha S. "Mining Knowledge of the Patient Record: "The Bayesian Classification to Predict and Detect Anomalies in Breast Cancer" *The Electronic Journal of Knowledge Management Volume 14 Issue 3 (pp128-139)* available online at www.ejkm.com

The application of information mining techniques to the medical domain are useful in extracting medical knowledge for diagnosis, decision-making, screening, monitoring, therapy support and patient management record. This will enhance patient safety and structure data during the acquisition process of data.

Text mining aims to extract useful knowledge from textual data or documents (Hearst, 1999), (Chen, 2001). Whereas, text mining is often considered a subfield of data mining, some text mining techniques have originated from other disciplines, such as information retrieval, information visualization, computational linguistics, and information science, (Chen, Fuller, Friedman and Hersh, 2005).

Text mining also has been applied to patient records and other clinical documents to facilitate knowledge management. It adopts a process similar to that of text mining from literature. For example, the system reported by Harris et al. (2003) extracts terms from clinical texts. Using natural language processing techniques, the MedLEE system (Friedman and Hripcsak, 1998) has been applied to free-text patient records. It extracts useful entities in order to identify patients having tuberculosis or breast cancer based on their admission chest radiographs and mammogram reports, respectively (Knirsch et al., 1996), (Jain and Friedman, 1997). Chapman et al. (2004) use a similar text mining approach for automated fever detection from clinical records to detect possible infectious disease outbreaks, from ((Chen, Fuller, Friedman and Hersh, 2005)).

Because of their predictive power, data mining techniques have been widely used in diagnostic and health care applications. Data mining is also used to extract rules from health care data. It has been used to extract diagnostic rules from breast cancer data (Kovalerchuk et al., 2001). The rules generated are similar to those created manually in expert systems and therefore can be easily validated by domain experts. Data mining has also been applied to clinical databases to identify new medical knowledge (Prather et al., 1997), (Hripcsak et al., 2002).

“Classification” is the most frequently used data mining function with a predominance of the implementation of Bayesian classifiers, neural networks, and SVMs (Support Vector Machines). Classification techniques are also applied to analyse various signals and their relationships with particular diseases or symptoms (Demigha$_b$, 2015).

Mining of electronic health records (EHRs) has the potential for establishing new patient stratification principles and for revealing unknown disease correlations (Jensen et al., 2012). An electronic health record (EHR), or electronic medical record (EMR), is a systematic collection of electronic health information about an individual patient or population. It is a record in digital format that is theoretically capable of being shared across different health care settings. EHRs may include a range of data, including demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics like age and weight, and billing information, (Demigha$_c$, 2015).

Digitized health information systems are expected to improve efficiency and quality of care and, reduce costs. Research efforts on how to analyse large amounts of patient data have been realized, however, they are still faced with problems of integrating scattered, heterogeneous data, in addition to ethical and legal obstacles that limit access to the data. It would be advisable that large-scale adoption of health information technology (HIT) infrastructure in the form of electronic health records (EHRs) and agreed standards for interoperability and schemes for privacy and consent, will improve this situation (Jensen et al., 2012).

In medicine and radiology Data Mining technology utilizes the available data in Radiology Information System (RIS) and Hospital Information System (HIS). It provides meaningful information adding value to diagnosis, plans further patient management, saves time, and reduces costs for the healthcare industry. By using the data analysis tool, the radiologist can make informed decisions and predict the future outcome of a particular imaging finding (Howell, 2012).

The application of data mining in imagery allows to obtain additional knowledge about specific features of different classes and the way in which they are expressed in the image (can help to find some inherent non-evident links between classes and their imaging in the picture). It can help to get some nontrivial conclusions and predictions can be made on the base of image analysis (Perner, 2000).

In this paper, we investigate the application of Data Mining (DM) techniques for mining knowledge of the patient record. We focus more specifically on available solutions offered by DM models and tools for improving the diagnosis and thus, the follow-up of patient. We provide in detail the acquisition process of DM in the medical field through the electronic patient record. We propose a methodology for mining medical knowledge based on the Bayesian

Classification. We use the Naïve Bayes Algorithm to develop this methodology. We illustrate the knowledge mining process by real examples from the breast cancer domain. We demonstrate through these illustrations how knowledge is better mined and thus, reused when applying concepts and techniques of Data Mining.

Section 2 defines and positions the Electronic Patient Record (EPR) in the literature and its combination with Data Mining.

2. The Electronic Patient Record (EPR) and Data Mining

The Electronic Patient Record (EPR) or Electronic Medical Record (EMR) system aims to represent data that accurately captures the state of the patient at all phases of examinations. Electronic Medical Records can improve patient safety and healthcare cost efficiency, but that depends on meaningful use of the data. This will require effective clinical decision support (CDS) content, particularly to drive clinical orders (labs, imaging, medications, etc.), the concrete manifestation of clinical decision making (Chen et al., 2014).

The research challenge is to transform data collected within EPRs into useful information. An EPR will be the first step. In addition, data must be leveraged through technology to inform clinical practice and decision-making. Without additional technology, EPR's are essentially just copies of paper-based records stored in electronic form. Modelling can be used to support clinical decisions that will provide a flexible, adaptable IT (Information Technology) framework which will be able to consolidate data from different sources. Data warehousing provides such an infrastructure (Inmon, 1996). As opposed to the EPR, a data warehouse does not have to be linked to a single provider organization, increasing its power, scope and utility (Bennett et al., 2010).

Soren Brunak (2012) notes that "the patient record becomes as information-rich as possible" and thereby "maximizes the data mining opportunities. Hence, electronic patient records further expand the possibilities regarding medical data mining thereby, opening the door to a vast source of medical data analysis." (Brunak, 2012).

Section 3 describes the DM process.

3. The Data Mining process

In knowledge management process, data mining technique can be used to extract and discover the valuable and meaningful knowledge from a large amount of data.

Data mining or "Knowledge Discovery in Databases" or KDD, (Fayyad, Piatetsky-Shapiro and Smyth, 1996) is an iterative process consisting of *data cleaning*, to remove noisy and inconsistent data, *data integration*, to combine multiple heterogeneous or homogeneous data sources, *data selection*, to consider only data relevant to the task and *data transformation* where data is transformed into forms appropriate for mining functions such as aggregation or summarization. Then data mining algorithms are employed to extract interesting and meaningful patterns from the data and present the knowledge to the domain expert in an informative manner. Figure 1 illustrates the Data Mining process.

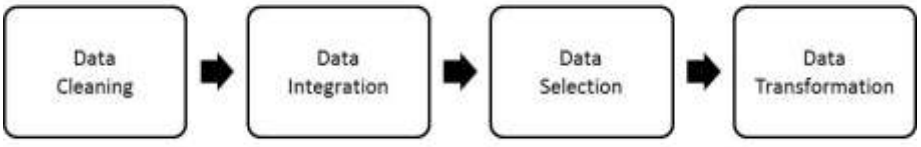

Figure 1: The Data Mining process

Section 4 describes Data Mining functionalities.

4. Data Mining Functionalities

The major tasks in data mining are classification and prediction; concept description; rule association; cluster analysis; outlier analysis; trend and evaluation analysis; statistical analysis and others. Classification and prediction tasks are among the popular tasks in data mining; and widely used in many areas especially for trend analysis and future planning.

4.1 Classification and prediction

Classification and prediction are forms of data analysis used to extract models to describe important data classes or to predict future data trends (Han et al., 2006).

Techniques used for data classification are decision tree, Bayesian methods, Bayesian network, rule-based algorithms, neural network, support vector machine, association rule mining, k-nearest-neighbour, case-based reasoning, genetic algorithms, rough sets, and fuzzy logic.

The classification process has two phases:

1. The learning process: the training data will be analysed by the *classification algorithms*. The *learned model* or *classifier* is represented in the form of *classification rules*
2. The classification process: the test data are used to estimate the accuracy of *the classification model* or *classifier*. If the accuracy is considered acceptable, the rules can be applied to the classification of new data

Figure 2 illustrates the process of classification and prediction in Data Mining.

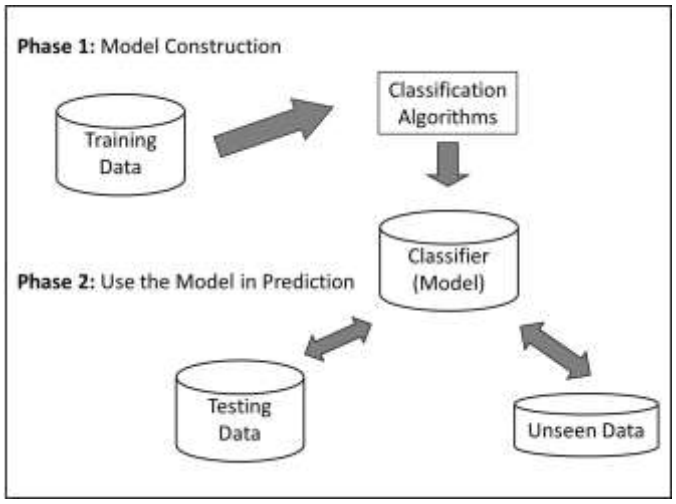

Figure 2: Classification and Prediction in Data Mining

Classification is one of the most common learning in data mining. This task aims at mapping a data item into one of several predefined classes. Examples of classification methods used as part of knowledge management include the classifying of the patients from primary health-care centers to specialists; the combination of the data mining and decision support approaches in planning of the regional health-care system; and the implementation of visualization method to facilitate KM and decision making processes, (Lavrac, Bohanec, Pur, Cestnik, Debeljak and Kobler, 2007).

4.2 Data Mining Tasks

Data mining tasks are grouping in Figure 3. We have detailed in (Demigha$_b$, 2015) data mining types with examples in radiology.

- Clustering: they consist of discovering groups and structures, without using known structures in the data. In medicine or radiology, data items may be grouped according to logical relationships or physician and radiologist affinities or preferences
- Association: data can be mined to identify associations. It is used in finding patterns of association among the attributes or variables and observations. In medicine variables may represent demographics and observations may represent observations. From the association technique, patterns that are discovered can be used to associate the student's profile for the most appropriate session/scenario, associated with trainee's attitude to performance
- Prediction and Classification: in this task the pattern discovered can be used to predict the percentage accuracy in trainee's (junior physician or junior radiologist) performance, behaviour, and attitudes, predict the performance

progress throughout the performance period, and also identify the best profile for different trainees in learning or training phase

Figure 3: Data Mining Tasks

Section 5 describes our methodology based on Data Mining concepts and techniques to manage data and knowledge deriving from the EPR.

5. A Methodology for Mining the knowledge EPR

We propose a classifier approach for detection of breast cancer disease and show how Naïve Bayes can be used for classification purpose. Due to the complexity of medical data, it will be better in certain projects or diagnoses to adapt existing algorithms or optimize their use to obtain better results (Iavindrasana et al., 2009). According to Harper (2005), "the best performing algorithm depends on the features of the data at hand as well as any preference of the end-user." (Harper, 2005).

The Bayesian Classification (supervised learning) method is based on the Naïve Bayes algorithm which has highly improved the screening operation in breast cancer. The Bayes algorithm is used to create models with predictive capabilities. It provides new ways of exploring and understanding data (Patil, 2014). We have used data extracted from real clinical cases taken from Electronic Patient Records (EPRs) specialized cancer radiology department. Relevant data extracted from medical (clinical or radiological) cases are mainly based on patient medical history and diagnosis. They are analysed, and classified in order to improve the breast screening management. We have experimented the Naive Bayes Classifier to the well-known *Wisconsin Prognostic Breast Cancer (WPBC) dataset* (UCI Machine Learning repository: http://archive.ics. uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/). The input features contain 10 relevant attributes.

5.1 Breast cancer screening

Breast cancer is considered as the second most common cancer in woman. The acquisition process of data deriving from screening operations is very complex and necessitates the creation of adequate models and requirement engineering tools. In senology (breast cancer radiology), BI-RADS (Breast Imaging and Reporting Data System) is used for the interpretation of patient diagnosis. It provides a standardized classification for mammographic studies. The BI-RADS system can inform radiologists about key findings, identify appropriate follow-up and management, and encourage the provision of educational and emotional support to patients. BI-RADS is classified into 7 categories: BI-RADS0: Incomplete, BI-RADS1: Negative, BI-RADS2: Benign finding, BI-RADS3: Probably benign, BI-RADS4: Suspicious abnormality, BI-RADS5: Highly suspicious of malignancy and BI-RADS6: Known biopsy with proven malignancy.

Breast cancer screening is based on the association of clinical data and mammographic image interpretation to conclude to the possible presence of abnormalities and their grade (i.e. benign, malignant, suspicious or doubtful). For a complementary diagnosis other imaging modalities can be used such as, Magnetic Resonance Imaging (MRI), Ultrasound and Tomography.

5.2 Classifying and detecting anomalies in breast cancer

Many methods have been developed to classify and detect anomalies in medical images. They are mainly based on feature's extraction using image-processing techniques (Antonie et al., 2001). These methods have proven their efficiency to assist physicians and radiologists for diagnosis and follow-up of patients.

Classification predicts categorical class labels, classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data. *Prediction* predicts unknown or missing values thanks to models continuous-valued functions.

Naive Bayes Classifier is a probabilistic classifier based on *the Bayes' theorem*, considering a strong Naïve independence assumption. A *Naive Bayes Classifier* considers that all attributes (features) independently contribute to the probability of a certain decision. Taking into account the nature of the underlying probability model, the *Naive Bayes Classifier* can be trained very efficiently in a supervised learning setting, working much better in many complex real-world situations, especially in the computer-aided diagnosis than one might expect (Belciug et al., 2008), (Gorunescu, 2006).

The proposed method classifies the diagnosis in 7 categories according to BI-RADS. Many researches using medical images classification have been realized of an automated image categorization. However, there is still no widely used method to classifying medical images (Antonie et al., 2001). In Antony et al., (2001), we found some reviews on developed methods for classifying and detecting anomalies in medical images, wavelets (Chen et al., 1997), (Wang et al., 1998), fractal theory (Li et al., 1997), statistical methods (Chan et al., 1998) and most of them used features extracted using image-processing techniques (Lai et al., 1989). In addition, some other methods were presented in the literature based on fuzzy set theory (Brazokovic et al., 1993), Markov models (Li et al., 1995) and neural networks (Dhawan et al., 1995), (Christoyianni et al., 2000). Most of the computer-aided methods proved to be powerful tools that could assist medical staff in hospitals and lead to better results in diagnosing a patient.

5.3 Bayesian Classification

A Bayesian Classification is based on:

- *Bayesian Classifier*: based on a statistical classifier which predicts class membership probabilities
- *Bayes Theorem* which estimate posterior probability
- *Naïve Bayesian classifier* which includes both a simple classifier that assumes attribute independence and a high speed when applied to large databases

5.3.1 Training phase

The *training data* are accompanied by labels indicating the class of the observations. New data is classified based on the *training set* and each *tuple/sample* is assumed to belong to a predefined class, as determined by the *class label attribute*. The *supervised classification* classifies data (constructs a model) based on the *training set* and the values (class labels) in a classifying attribute and uses it in classifying new data. Figure 4 illustrates the process of *Classification*.

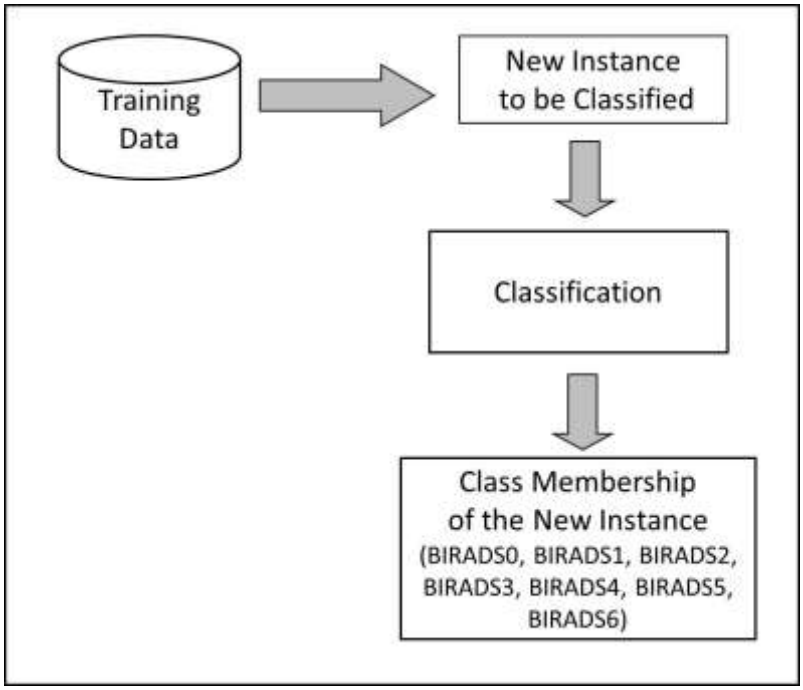

Figure 4: Classification

5.3.2 Testing phase

Testing phase consists of estimating the accuracy of the model i.e. unknown or missing values. Accuracy rate is the percentage of test set samples that are correctly classified by the model.

The models continuous-valued functions predict unknown or missing values. Figure 5 illustrates the process of *Prediction*.

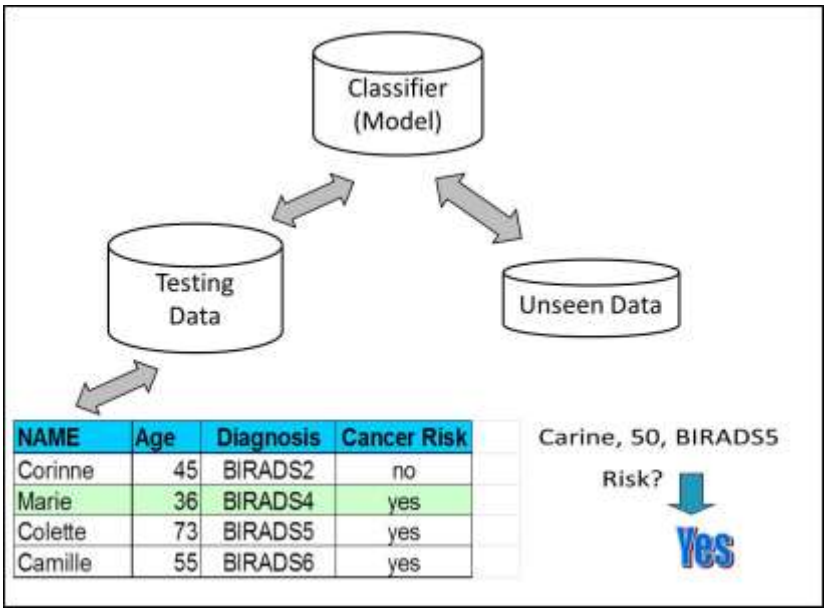

Figure 5: Prediction

5.3.3 Naïve Bayes

Naive Bayes is the simplest of the classification algorithms. It builds patterns by counting the correlations between all different states of the input attributes and all different states of the output attributes. Attributes can only have discrete values. Naive Bayes is based on Bayes' theorems and is naive in the sense that it does not take possible

dependencies among the input attributes into account. Naive Bayes is used in the beginning of the data mining process to quickly explore the data but can also be a powerful predictor in some situations.

Patterns generated by *Naive Bayes* include so called attribute characteristics which can be interpreted as the main influencers. Attribute characteristics are expressed as an attribute-state combination with an associated frequency which indicates the proportion of cases with the target output state that also had this specific input attribute-state combination.

Theorem

- Let X be a data sample whose class label is unknown
- Let Hi be the hypothesis that X belongs to a particular class Ci
- P(Hi) is class prior probability that X belongs to a particular class Ci
  Can be estimated by ni/n from training data samples

  n is the total number of training data samples

  ni is the number of training data samples of class Ci

$$P(H_i|X)=\frac{P(X|H_i)P(H_i)}{P(X)} \quad (1)$$

Algorithm

The Na*ïve* Bayes algorithm is based on the Bayesian theorem as given by equation (1).

Steps in algorithm are as follows:

Step1. Let D be a data sample is represented by an n dimensional feature vector, X = (x1, x2….. xn), depicting n measurements made on the sample from n attributes, respectively A1, A2, An.

Step2. Suppose that there are m classes, C1, C2……Cm. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned if and only if:

P(Ci/X) > P(Cj/X) for all 1< = j< = m and j ≠ i

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes theorem.

Step3. As P(X) is constant for all classes, only P(X|Ci)*P(Ci) need be maximized. If the class prior probabilities are not known, then it is often assumed that the classes are equally likely, that is, P(C1) = P(C2) = …..= P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci) = P (Ci) =|Ci, D|/|D|, where |Ci, D| is the number of training samples of class Ci in D.

5.4 Naïve Bayes Classifier in Breast Cancer Prediction

An advantage of the *Naive Bayes Classifier* is that it requires a small amount of *training data* to estimate the parameters (means and variances of the variables) necessary for classification. It performs better in many complex real world situations like Spam Classification, Medical Diagnosis, and Weather forecasting. It is suited when dimensionality of input is high (Kharya and Soni, 2016).

Classification aims to build the *Classifier Model* based on *Bayes theorem*. Then, the classifier is used to predict the group attributes of new cases from domain based on the values of attributes. This prediction technique assigns patients to either a “benign“ group (BI-RADS1, BI-RADS2, BI-RADS3) that is non-cancerous or a “malignant” group (BI-RADS4, BI-RADS5, BI-RADS6) that is cancerous.

5.5 Weighted Naive Bayesian in Breast Cancer Prediction

The *Weighted Naive Bayes Classifier (WNBC)* assigns different weights to different attributes according to their predicting capabilities by consulting the domain expert physicians. Domain based weights are used to assign weight of each attribute using expert knowledge, (Kharya and Soni, 2016).

5.6 Results

We investigate the potential contribution of the *Naive Bayesian Classification* methodology as a reliable support in computer-aided diagnosis of such events, using the well-known *Wisconsin Prognostic Breast Cancer* dataset. For training the system Wisconsin Datasets consisting of 699 records with 9 medical attributes with 2 class labels have been used i.e. benign and malignant. Dataset is available in (Wolbergand and Mangasarian, 1990). The testing diagnosing accuracy, that is the main performance measure of the classifier, was about 91.50%, in accordance with the performance of other wellknown Machine Learning techniques. Table 1 groups name attributes, range of values and weights selected from the Wisconsin Dataset.

Table 1: Normalized Breast. D20.N699.C2 dataset with range of values and weights.

S. No	Attribute Name	Domain	Range	Weight S (1 to 10)
1	Clump Thickness	1-10	1	4
			2	9
2	Uniformity of cell size	1-10	3	5
			4	9
3	Uniformity of cell shape	1-10	5	4
			6	9
4	Marginal Adhesion	1-10	7	5
			8	9
5	Single Epithelial cell size	1-10	9	3
			10	9
6	Bare Nuclei	1-10	11	4
			12	8
7	Bland Chromatin	1-10	13	4
			14	9
8	Normal Nucleoli	1-10	15	5
			16	9
9	Mitoses	1-10	17	3
			18	9
10	Output (class label representing 2 type of breast cancer class)	2 for benign and 4 for malignant	19/20	

Based on the *Weighted Count* and *Weighted Probability* definitions 1 and 2 illustrated by equations 2 and 3. In (Kharya

and Soni, 2016), the classifier model is built with all 699 records and testing is also performed on 699 records of the Wisconsin Dataset, (Wolbergand and Mangasarian, 1990).

Definition 1. Attribute Weight Attribute weight is assigned depending upon the domain. Weight of different attribute in predicting the probability of breast cancer is given in Table 1.

*Definition 2***.** Weighted Count

$$\text{Sum of Records weight having condition attribute column} = \frac{\text{“Class Value1(S)” given Class Label=S}}{\text{Total Record Weights}} \quad (2)$$

Definition 3. Weighted Probability

$$\frac{\text{Weight Count of particular attribute=“Yes”/”No”}}{\text{Weight Count of total “Yes”/”No”}} \quad (3)$$

On building the Classifier Model for Breast Cancer Prediction based on Weighted Naive Bayes Classifier, the accuracy is found to be 92% for WNBC.

A comparison of the classification results of different classifiers found in the literature, indicates that introducing the novel concept of Weights on Naive Bayes Classifier Breast cancer detection provides very promising results with strong and efficient predictive results which outperforms others classifiers and the classification results were consistent with some of the highest results, (Kharya and Soni, 2016).

Section 6 provides a discussion and a conclusion of the overall work presented in the paper.

6. Discussion and conclusion

This paper has presented a methodology for mining knowledge of the patient record based on a Bayesian Classification combined to the Naïve Bayes Algorithm. The Bayesian Classification for breast cancer was used to determine the higher prediction accuracy.

We have put in value the necessity to mine the Electronic Patient Record. We have reviewed some scientific and practical papers in this problematic. The Bayesian Classification has many advantages in the medical field. Thanks to the models of classification and prediction, we have extracted the hidden knowledge from an ‘experience base’ of senology (breast cancer) and on the other hand predicted patients with their breast disease. These models answer to complex queries through a simple way and allow an easy access to information and accuracy. We have illustrated the validity of them by 100 cases of patients affected by the breast cancer disease. It is essential to properly collect and prepare the data, and to check the models against the real world.

We have used the Breast Imaging and Reporting Data System for categorizing the vocabulary which is a standard vocabulary and helped us in applying the Bayesian Classification.

The methodology we have developed enables exploration and analysis, by automatic means, of large quantities of data related to breast cancer characteristics, in order to obtain an optimal prediction of recurrent events. The purpose is to demonstrate the suitability and ability of the Naive Bayes methodology in Classification/Prediction problems in breast cancer screening and diagnosis.

We have investigated the potential contribution of the Naive Bayesian Classification methodology as a reliable support in computer-aided diagnosis using the well-known Wisconsin Prognostic Breast Cancer dataset. Finally, we have demonstrated the suitability and ability of the Naive Bayes methodology in Classification/Prediction problems in breast cancer using the Weighted Naive Bayes Classifier.

The potential of Data Mining techniques in medical field allows to improve the quality and decrease the cost. They are very helpful in extracting medical knowledge for diagnosis, decision-making, screening, monitoring, therapy support and patient management.

References

ACR/NEMA [online] BI-RADS, ACR, American College of Radiology, http://www.acr.org/.

Antonie, M.L., Zaıane, O.R. and Coman, A. (2001) "Application of Data Mining Techniques for Medical Image Classification," *In Proceedings of the Second International Workshop on Multimedia Data Mining,* San Francisco, USA.

Belciug, S. (2008) "Bayesian classification vs. k-nearest neighbor classification for the non-invasive hepatic cancer detection," *8th International Conference on Artificial Intelligence and Digital Communications,* Craiova, (Research Notes in Artificial Intelligence and Digital Communications), pp 31–35.

Bennett, C. C., and Doub, T.W. (2010) "Data mining and electronic health records: Selecting optimal clinical treatments in practice," *Proceedings of the 6th International Conference on Data Mining*, pp 313–318.

Brazokovic, D. and Neskovic, M. (1993) "Mammogram screening using multiresolution-based image segmentation," *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 7, No. 6, pp 1437–1460.

Brunak, S. (2012) "Analyzing Medical Data," *Communications of the ACM*, Vol. 55, No. 6, pp 13-15.

Chan, H. et al. (2007) "Computerized analysis of mammographic microcalcifications," *In morphological and feature spaces,* Medical Physics, Vol. 25, No. 10, pp 2007–2019.

Chapman, W. W., Dowling, J. N., and Wagner, M. M. (2004) "Fever Detection from Free-text Clinical Records for Biosurveillance," *Journal of Biomedical Informatics*, Vol. 37, pp 120-127.

Chen, H. (2001) "*Knowledge Management Systems: A Text Mining Perspective*," Tucson, AZ: The University of Arizona.

Chen, J.H. and Altman, R.B. (2014) "Automated Physician Order Recommendations and Outcome Predictions by Chen. K. K et al. (2007) "Constructing a Web-based Employee Training Expert System with Data Mining Approach," the 9th IEEE International Conference on E-Commerce Technology and The 4th IEEE International Conference on Enterprise Computing, E-Commerce and Eservices (CEC-EEE). Data-Mining Electronic Medical Records*," Proceedings of the AMIA Clinical Research Informatics.*

Chen, H. and Chau, M. (2004) "Web Mining: Machine Learning for Web Applications," *Annual Review of Information Science and Technology*, Vol. 38, pp 289-329.

Chen, C. and Lee, G. (1997) "Image segmentation using multiresolution wavelet analysis and expectation maximization (em) algorithm for digital mammography*," International journal of imaging systems and Technology,* Vol. 8, No. 5, pp 491–504.

Chen, H, Fuller. S., Friedman. C. and Hersh. W. (2005) "Knowledge Management, Data Mining, and Text Mining in Medical Informatics," Vol. 8, pp 3-33, eds. Chen, H, Fuller. S., Friedman. C., Hersh. W.

Chien, C. F. and Chen, L. F. (2007) "Using Rough Set Theory to Recruit and Retain High- Potential Talents for Semiconductor Manufacturing*," IEEE Transactions on Semiconductor Manufacturing*, Vol. 20, No 4, pp 528–541.

Chien, C. F. and Chen, L. F. (2008) "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry," *Expert Systems and Applications, Vol*. 34, No 1, pp 380–290.

Christoyianni, I. et al. (2000) "Fast detection of masses," In computer-aided mammography, *IEEE Signal Processing Magazine,* pp 54–64.

Demigha, S. (2009) "A Data Warehouse System to Help Assist Breast Cancer Screening in Diagnosis, Education and Research," *In CSA the Second International Conference on Computer Science and its Applications, IEEE*, Jeju, Korea (South), pp 1–6.

Demigha, S. (2012) "A Business Process Multidimensional Modeling with the Case-Based Reasoning in Breast Radiology," *In the World Congress in Computer Science, Computer Engineering, and Applied Computing – the International Conference on e-Learning, e-Business, Enterprise Information System, and e-Government, EEE,* pp 473–477, Las Vegas, NV (USA).

Demigha$_a$, S. (2015) "Knowledge Management and Intellectual Capital in an Enterprise Information System", *ECKM 16th European Conference on Knowledge Management*, pp 213-221, 3-4 September, Udine, Italy.

Demigha$_b$, S. (2015) "Data Mining for Breast Cancer Screening," *In the 10th IEEE International Conference on Computer Science & Education, IEEE ICCSE*, pp 59–63, Fitzwilliam College, Cambridge University, Cambridge, UK.

Demigha$_c$, S. (2015) "Mining Knowledge of the Patient Record**,"** *The 12th International Conference on Intellectual Capital, Knowledge Management and Organisational Learning, ICICKM***,** Bangkok University Bangkok, pp 71-79, Thailand.

Dhawan, A. et al. (1995) "Radial-basis-function-based classification of mammographic microcalcifications using texture features," *In Proc. of the 17th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vol. 1, pp 535–536.

Dumitru, D. (2009) "Prediction of recurrent events in breast cancer using the Naive Bayesian classification," *Annals of University of Craiova, Math. Comp. Sci. Ser*. Vol. 36, No. 2, pp 92-96, ISSN: 1223–6934.

Dunham, M. H. (2002) "Data Mining: Introductory and Advanced Topics," New Jersey, USA: Prentice Hall.

Fayyad, U., Shapiro, G.P. and Smyth, P. (1996) "Knowledge Discovery and Data Mining: Towards a Unifying Framework," *In Proc. 2nd International Conference on Knowledge Discovery and Data Mining. AAAI Press*, pp 82–8.

Fichman, R.G., Kohli, R. and Krishnan, R. (2011) "The Role of Information Systems in Healthcare: Current Research and Future Trends," *Information Systems Research*, Vol. 22, No. 3, pp 419–428.

Friedman, C. and Hripcsak, G. (1998) "Evaluating Natural Language Processors in the Clinical Domain," *Methods of Information in Medicine*, Vol. 37, pp 334-344.

Gorunescu, F. (2006) "Data Mining: Concepts, models and techniques," *Blue Publishing House*, Cluj-Napoca.

Han, J. and Kamber, M. (2006) "Data Mining: concepts and techniques," *Second Edition, Editor, Morgan Kaufmann.*

Harris, M. R., Savova, G. K., Johnson, T. M., and Chute, C. G. (2003) "A Term Extraction Tool for Expanding Content in the Domain of Functioning, Disability, and Health: Proof of Concept," *Journal of Biomedical Informatics*, Vol. 36, pp 250-259.

Harper, P.R. (2005) "A review and comparison of classification algorithms for medical decision making," Health Policy, Vol. 71, pp 315-31.

Hearst, M. A. (1999) “Untangling Text Data Mining,” In *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, June 20-26.

Hripcsak, G., Austin, J. H., Alderson, P. O., and Friedman, C. (2002) “Use of Natural Language Processing to Translate Clinical Information from a Database of 889,921 Chest Radiographic Reports,” *Radiology*, Vol. 224, No.1, pp 157-163.

Howell, W.L. (2012) “Data Mining and Analytics in Radiology,” *PACS and Informatics, RIS.*

Inmon, J. W.H. (2005) “Building the Data Warehouse,” *Fourth Edition, Wiley Publishing in Indianapolis,* Indiana.

Iavindrasana, J. et al. (2009) “Clinical Data Mining: a Review,” Geissbuhler A, Kulikowski C, editors. *IMIA Yearbook of Medical Informatics*.

Jain, N. L. and Friedman, C. (1997) “Identification of Findings Suspicious for Breast Cancer Based on Natural Language Processing of Mammogram Reports,” In *Proceedings of the Fall AMIA Conference*, Philadelphia, USA, pp 829-833.

Jensen, P. B., Jensen,. L. J. and Brunak,. *S*. (2012) “Mining electronic health records: towards better research applications and clinical care,” *Macmillan Publishers Limited*, Vol. 13, pp 395–405.

Kharya.S and Soni.S, (2016) “Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection,” International Journal of Computer Applications (0975 – 8887) Vol. 133, No.9.

Kovalerchuk, B., Vityaev, E., and Ruiz, J. F. (2001) “Consistent and Complete Data and ‘Expert’ Mining in Medicine,” In Cios, K. J. (Ed.), *Medical Data Mining and Knowledge Discovery*, New York, USA: Physica-Verlag.

Knirsch, C.A., Jain, N. L., Pablos-Mendez, A., Friedman, C., and Hripcsak, G. (1996). “Respiratory Isolation of Tuberculosis Patients Using Clinical Guidelines and an Automated Clinical Decision Support System,” *Infection Control and Hospital Epidemiology*, Vol.19, No. 2, pp 94-100.

Lai, S., et al. (1989) “On techniques for detecting circumscribed masses in mammograms*.” IEEE Trans. Medical Imaging,* Vol. 8, No. 4, pp 377–386.

Lavrac. N, Bohanec. M, Pur. A, Cestnik. B, Debeljak. M, Kobler. A. (2007) “Data mining and visualization for decision support and modeling of public health-care resources,” *J Biomed Inform.* Vol. 40, No. 4, pp 438-47.

Li, H. et al. (1995) “Marcov random field for tumor detection in digital mammography,” *IEEE Trans. Medical Imaging,* Vol. 14, No. 3, pp 565–576.

Li, H., et al. (1997) “Fractal modeling and segmentation for the enhancement of microcalcifications in digital mammograms,” *IEEE Trans. Medical Imaging*, Vol. 16, No. 6, pp 785–798.

Medhekar, D. S., Bote, M. P. and Deshmuk, S.D. (2013) “Heart Disease Prediction System using Naive Bayes,” International journal of enhanced research in science technology & engineering, Vol. 2, No. 3.

Patil. R. (2014). “Heart Disease Prediction System using Naïve Bayes and Jelinek-mercer smoothing,” *International Journal of Advanced Research in Computer and Communication Engineering,* Vol. 3, No. 5.

Perner, P. (2000) “Mining Knowledge in Medical Image Databases,” *In Data Mining and Knowledge Discovery: Theory, Tools, and Technology,* Belur V. Dasarathy (eds.), *Proceedings of SPIE*, Vol. 4057, pp 359–369.

Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., and Hammond, W. E. (1997) “Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse,” In *Proceedings of the AMIA Annual Symposium Fall*, pp 101-105.

Radhika. (2008) “Topics in Biomedical Informatics Data Mining and its applications and usage in medicine,” pp 1–22.

Ranjan, J. (2008) “Data Mining Techniques for better decisions in Human Resource Management Systems,” *International Journal of Business Information Systems*, Vol. 3, No. 5, pp 464–481.

Wang, T. and Karayiannis, N. (1998) “Detection of microcalcifications,” *In digital mammograms using wavelets. IEEE Trans. Medical Imaging,* Vol.17, No. 4, pp 498–509.

Wang, H. & Wang, S. (2008) “A knowledge management approach to data mining process for business intelligence. Industrial Management & Data Systems,” Vol. 108, No. 5, pp 622-634.

William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", *Proceedings of the National Academy of Sciences*, U.S.A., Vol. 87, pp 9193-9196.