

## Analysis On Classification Techniques In Mammographic Mass Data Set

Mrs. K. K. Kavitha \*, Dr. A Kangaialmmal\*\*, Mr. K. Satheesh\*\*\*

\*(Department of Computer Science, Selvamm Arts & Science College, Namakkal-03)

\*\* (Department of Computer Science, Government Arts College, Salem-07)

\*\*\*(Department of Computer Science, Selvamm Arts & Science College, Namakkal-03)

### ABSTRACT

Data mining, the extraction of hidden information from large databases, is to predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data-Mining classification techniques deals with determining to which group each data instances are associated with. It can deal with a wide variety of data so that large amount of data can be involved in processing. This paper deals with analysis on various data mining classification techniques such as Decision Tree Induction, Naïve Bayes , k-Nearest Neighbour (KNN) classifiers in mammographic mass dataset.

**Keywords** - Classification , Data Mining, Decision Tree , Mammographic Dataset, Naïve Bayes Classifier.

### I. INTRODUCTION

The field of data mining is an emerging research area with important applications in Engineering, Science, Medicine, Business and Education[9]. The size of data base in educational application is large where the number of records in a data set can vary from some thousand to thousand of millions. The size of data is accumulated from different fields exponentially increasing. Data mining has been used in different methods at the intersection of Machine Learning, Artificial Intelligence, Statistics and Database Systems.

Cancer begins in cells, the building blocks that make up tissues. Tissues make up the breasts and other parts of the body. Normal cells grow and divide to form new cells as the body needs them. When normal cells grow old or get damaged, they die, and new cells take their place. Sometimes, this process leads to wrong results. New cells form when the body does not need them, and old or damaged cells do not die as they should. The built-up of extra cells often forms a mass of tissue called a lump, growth, or tumor. Cancer that forms in the tissues of breast, usually in the ducts (tubes that carry milk to the nipple) and in the lobules (glands that make milk) is the breast cancer [1]. It occurs in both men and women; although in male, breast cancer is very rare. Breast cancer is one of the leading cancers in the female population. About 25% of all cancers diagnosed in women are breast cancers and about 20% of all lethal cancers are breast cancers. Mammography has been proven to be the most reliable method and it is the key screening tool for the early detection of breast cancer. Mammography

is highly accurate. On average, mammography will detect about 80–90% of the breast cancers in women without any recognizable symptoms.

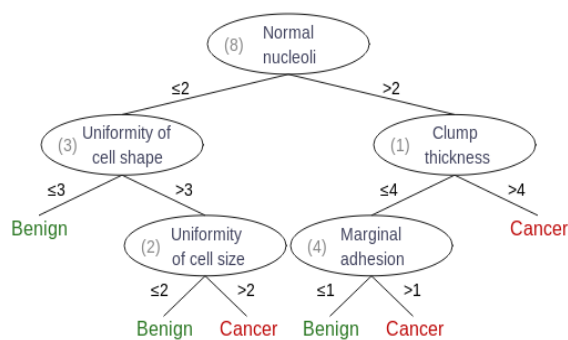
Classification technique is capable of processing a wider variety of data and is growing in popularity. This paper deals with analysis on various data mining classification techniques in mammographic mass datasets.

The next section presented is Decision Tree Induction. Section II describes Naïve Bayesian Classification whereas k-nearest neighbor (KNN) classifier is described in section III. Finally, the last section concludes this work.

### II. Decision tree induction

Decision Tree Induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label[9],[4].

The feature that best divides the training data would be the root node of the tree. There are numerous methods for finding the feature that best divides the training data such as information gain (Hunt et al., 1966) and gini index (Breiman et al., 1984). While myopic measures estimate each attribute independently, Relief algorithm (Kononenko, 1994) estimates them in the context of other attributes. However, a majority of studies have concluded that there is no single best method (Murthy,1998).



**Figure 1: Decision Tree for Breast Cancer**

The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner [9]. Algorithm for decision tree is described in Fig. 2.

### Algorithm for Decision Tree Induction

- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a **top-down recursive divide-and-conquer manner**
  - At start, all the training examples are at the root
  - Attributes are categorical (if continuous-valued, they are discretized in advance)
  - Examples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., **information gain**)
- Conditions for stopping partitioning
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – **majority voting** is employed for classifying the leaf
  - There are no samples left

**Figure 2: Decision Tree Algorithm**

### III. NAÏVE BAYESIAN CLASSIFICATION

Naïve Bayes is a simple technique for constructing classifier models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable[13].

Abstractly, naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector  $\mathbf{x} = (x_1, \dots, x_n)$  representing some  $n$  features (dependent variables), it assigns to this instance probabilities for each of  $K$  possible outcomes or classes.

Using Bayes theorem, the conditional probability can be decomposed as:

$$p(C_k|\mathbf{x}) = \frac{p(C_k) p(\mathbf{x}|C_k)}{p(\mathbf{x})} \quad (1)$$

In general, using Bayesian probability terminology, the above equation can be written as:

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \quad (2)$$

Using the chain rule for repeated applications of the definition of conditional probability[11]:

$$p(C_k|x_1, \dots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^n p(x_i|C_k) \quad (3)$$

One can easily estimate the probabilities from the training tuples.

### IV. KNN CLASSIFICATION

K Nearest Neighbors (KNN) is a simple algorithm that stores all available cases and classifies new cases based on a specific similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition[10].

Nearest neighbor classifiers are based on learning by analogy. The training samples are described by  $n$  dimensional numeric attributes. Each sample represents a point in an  $n$ -dimensional space. Similarly, all of the training samples are stored in an  $n$ -dimensional pattern space. When given an unknown sample, a  $k$ -nearest neighbor classifier searches the pattern space for the  $k$ - training samples that are closest to the unknown sample.

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its  $K$ -nearest neighbors measured by a distance function. If  $K = 1$ , then the case is simply assigned to the class of its nearest neighbor[6]:

#### Distance functions

Euclidean	$\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$
Manhattan	$\sum_{i=1}^k  x_i - y_i $
Minkowski	$\left( \sum_{i=1}^k ( x_i - y_i ^q) \right)^{1/q}$

(4)

It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables the Hamming distance must be used. It also brings up the issue of

standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

### Hamming Distance

$$D_H = \sum_{i=1}^k |x_i - y_i| \quad (5)$$

The k-nearest neighbors algorithm is amongst the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k-nearest neighbors. k is a positive integer, typically small. If k= 1, then the object is simply assigned to the class of its nearest neighbor. In binary classification problems (two class), it is helpful to choose k to be an odd number as this avoids tie in votes.

The same method can be used for regression, by simply assigning the property value for the object to be the average of the values of its k-nearest neighbors. It can be useful to weigh the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

The neighbors are taken from a set of objects for which the correct classification is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. In order to identify neighbors, the objects are represented by position vectors in a multidimensional feature space. It is usual to use the Euclidian distance, though other distance measures, such as the Manhannttan distance could be used in principle instead. The k-nearest neighbor algorithm is sensitive to the local structure of the data.

## V. EXPERIMENTAL RESULTS

### 1. DATASET ACQUISITION

This breast cancer domain was obtained from the University Medical Centre, Institute of Oncology, and Ljubljana, Yugoslavia. This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.

### 2. TOOL

In this paper, we applied weka tool for finding classification accuracy. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and

visualization, etc.. It is also well-suited for developing new machine learning schemes[14].

## VI. PERFORMANCE EVALUATION

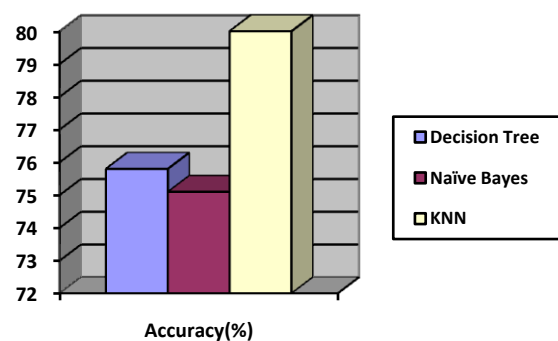
In this paper, we discussed about three major data mining classification algorithms. Now, we had given some results by applying these algorithms in the following table with some classification errors such as MAE, RMSE, RAE and so on.

**Table 1 Comparison of Classification Algorithms**

	Accuracy(%)	MAE	RMSE	RAE(%)
<b>Decision Tree</b>	75.8	0.3658	0.4269	87.44
<b>Naïve Bayes</b>	75.1	0.3012	0.4278	72.00
<b>KNN</b>	80	0.2679	0.3671	64.04

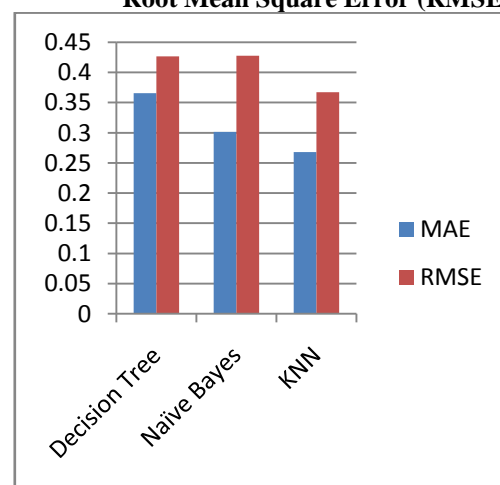
Now, we represent the comparison results in the following fig. 3 through 5.

### i) Accuracy:



**Figure 3: Accuracy**

### ii) Mean Absolute Error (MAE) and Root Mean Square Error (RMSE):



**Figure 4: MAE & RMSE**

iii) Relative Absolute Error:

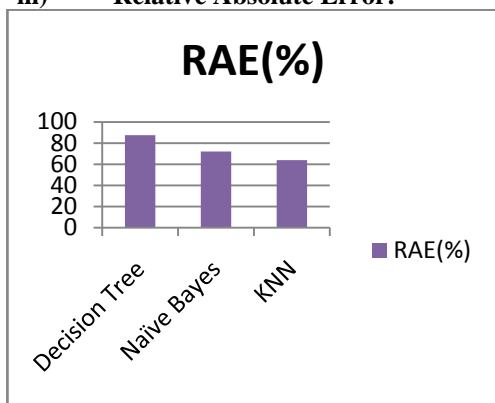


Figure 5: RAE

## VII. CONCLUSION

From the above analysis, it is concluded that KNN Classifier provides high accuracy than Decision Tree approach and the Naïve Bayes classifier. In future, this work can be extended for implementing these algorithms for uncertain datasets.

## REFERENCES

### Books:

- [1] Aley Abraham, Blessings of My Breast Cancer, *Innova Publishing*, 2012.
- [2] Arun K Pujari, Data mining Technique, Published by *Universities Press (I) Pvt. Ltd*, Hyderabad, India, 2001.
- [3] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, *Second Edition, Elsevier Inc.*, 2006.

### Journal Papers:

- [4] Bharati M. Ramageri, Data Mining Techniques and Application, *Indian Journal of Computer Science and Engineering*, Vol. 1 No. 4, pp. 301-305.
- [5] Jiazhen He, Yang Zhang, Xue Li & Peng Shi, Learning naive Bayes classifiers from positive and unlabelled examples with uncertainty, *International Journal of Systems Science*, Volume 43, Issue 10, 2012.
- [6] Kotsiantis. S. B., Supervised Machine Learning: A Review of Classification Techniques, *Informatica*, pp.249-268, 2007.
- [7] Narasimha Murty, M., Susheela Devi, V. Pattern Recognition: *An Algorithmic Approach, Springer Science & Business Media*, 2011.
- [8] Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, David Scuse, Weka Manual for Version 3-6-11, *University of Waikato, Hamilton, New Zealand*, 2014.

- [9] Sumit Garg, Arvind K. Sharma, Comparative Analysis of Data Mining Techniques on Educational Dataset, *International Journal of Computer Applications* (0975 – 8887) Volume 74, No.5, July 2013.
- [10] Suresh Chandra Satapathy, Siba K Udgata, Bhabendra Narayan Biswal, "Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)", *Springer*, 2013.

### Conference Proceedings:

- [11] Radovic, M. R&D Center for Bioeng., BioIRC, Kragujevac, Serbia Djokovic, M., Peulic, A., Filipovic, N., Application of data mining algorithms for mammogram classification, *Bioinformatics and Bioengineering (BIBE), IEEE 13th International Conference*, 2013.
- [12] Thair Nu Phyu, Survey of Classification Techniques in Data Mining, *Proceedings of the International Multi Conference of Engineers and Computer Scientists*, Vol I IMECS 2009, March 18 - 20, 2009.

### Web References:

- [13] [http://en.wikipedia.org/wiki/Naive\\_Bayes\\_classifier](http://en.wikipedia.org/wiki/Naive_Bayes_classifier).
- [14] <http://www.cs.waikato.ac.nz/ml/weka/>.