DETECTION OF DIABETES AND CHOLESTEROL THROUGH CLASSIFICATION ALGORITHMS

B.Iswarya¹, T.Sathyapriya²

¹Research scholar, Department of Computer Science, Theivanai Ammal College for Women, Villupuram.

²Assistant Professor, Department of Computer science, Theivanai Ammal College for Women, Villupuram.

Abstract: Classification is a machine learning system used to predict group membership for data instance. Classification is mainly used in medical field. Because medical dataset is difficult to compare with the other data mining techniques like Clustering, Association, and Regression. The previous research predicts only the sugar level using five classification algorithms like C4.5, Apriori, Naive Bayes, Support Vector Machine, and K-Nearest Neighbour algorithms. The proposed system detect both sugar and cholesterol level using three classification algorithms like Decision Tree, Naive Bayes, Apriori algorithms. The parameters of the diseases are considered for classification. Based on the classification result of the Decision tree, Naive Bayes, Apriori algorithms, the accuracy of sugar and cholesterol level will be provided.

Keywords - C4.5, SVM, K-NN, Apriori, Naive Bayes.

I. INTRODUCTION

Data mining refer to extracting or “Mining” knowledge from large amounts of data. Data mining also known as Knowledge-Discovery in Database (KDD) is the process of automatically searching large volumes of data for patterns. Data mining applies many older computational techniques form statistics, machine learning and pattern recognition.

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. Machine learning is a scientific discipline that deals with the construction and study of algorithms that can learn from data. Such algorithms operate by building a model based on input and using that to make predictions or decisions. Pattern recognition is a branch of machine learning that focuses on the recognition of patterns and regularities in data, although is in some cases considered to be nearly synonymous with machine learning.

Pattern recognition systems are in many cases trained from labelled “training” data (supervised learning), but when no labelled data are available other algorithms can be used to discover previously unknown patterns (unsupervised learning). The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: Descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions. Predictive techniques described three types are Classification, Regression, and Forecasting techniques. Descriptive techniques described three types are Clustering, Association, and Sequence Discovery.
II. CLASSIFICATION

In machine learning and statistics, the classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known. An example would be assigning a given email into "spam" or "non-spam" classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient gender, blood pressure, presence or absence of certain symptoms, etc.

In the terminology of machine learning, classification is considered an instance of supervised learning, the learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance.

Often, the individual observations are analyzed into a set of quantifiable properties, known variously explanatory variables, features, etc. These properties may variously be categorical (e.g. "A", "B", "AB" or "O", for blood type), ordinal (e.g. "large", "medium" or "small"), integer-valued (e.g. the number of occurrences of a part word in an email) or real-valued (e.g. a measurement of blood pressure). Other classifiers work by comparing observations to previous observations by means of a similarity or distance function.

An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category.

Terminology across fields is quite varied. In statistics, where classification is often done with logistic regression or a similar procedure, the properties of observations are termed explanatory variables, and the categories to be predicted are known as outcomes, which are considered to be possible values of the dependent variable. In machine learning, the observations are often known as instances, the explanatory variables are termed features (grouped into feature), and the possible categories to be predicted are classes.

III. ALGORITHMS USED

A. C4.5 Algorithm

Decision trees are controlling categorization algorithms. Accepted decision tree algorithms consist of C4.5. At the equivalent time as the name imply, this performance recursively separate inspection in branches to build tree for the purpose of improving the calculation accuracy. Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each belong-ing to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs.

C4.5 generates classifiers expressed as decision trees, but it can also construct classifiers in more comprehensible rule set form.

B. The K-Nearest Neighbour Algorithm

The k-nearest neighbour’s algorithm is a technique for classifying objects based on the next training data in the feature space. It is among simplest of all mechanism learning algorithms. The algorithm operates on a set of d-dimensional vectors, D = {x_i | i = 1...N}. 
where \( x_i \in k^d \) denotes the \( i \)th data point. The algorithm is initialized by selecting \( k \) points in \( k^d \) as the initial \( k \) cluster representatives or “centroids”. Techniques for selecting these primary seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data \( k \) times. Then the algorithm iterates between two steps till junction:

Step 1: Data Assignment each data point is assigned to its adjoining centroid, with ties broken arbitrarily. This results in a partitioning of the data.

Step 2: Relocation of “means”. Each group representative is relocating to the center (mean) of all data points assigned to it. If the data points come with a possibility measure (Weights), then the relocation is to the expectations (weighted mean) of the data partitions.

C. Naive Bayes Algorithm

Naïve Bayes Classifier is a term dealing with simple probabilistic classifier based on applying Bayes Theorem with strong independence assumptions. It assumes that the presence or absence of particular feature of a class is unrelated to the presence or absence of any other feature. The Naive Bayes algorithm is based on conditional probabilities.

It uses Bayes' theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Bayes ‘Theorem finds the probability of an event occurring given the probability of another event that has already occurred. If \( B \) represents the dependent event and \( A \) represents the prior event, Bayes' theorem can be stated as follows.

\[
\text{Probability (B given A)} = \frac{\text{Probability (A and B)}}{\text{Probability (A)}}
\]

To calculate the probability of \( B \) given \( A \), the algorithm counts the number of cases where \( A \) and \( B \) occur together and divides it by the number of cases where \( A \) occurs alone.

D. Support Vector Machines (SVM)

Support vector machines are a moderately new-fangled type of learning algorithm, originally introduced. Naturally, SVM aim at pointing for the hyper plane that most excellently separates the classes of data. SVMs have confirmed the capability not only to accurately separate entities into correct classes, but also to identify instances whose classification is not supported by data. Although SVM are comparatively insensitive define

Distribution of training examples of each class SVM can be simply extended to perform numerical calculations. Two such extension, the first is to extend SVM to execute regression analysis, where the goal is to produce a linear function that can fairly accurately approximate the target function. An extra extension is to learn to rank elements rather than produce a classification for individual elements. Ranking can be reduced to comparing pairs of instance and producing a +1 estimate if the pair is in the correct ranking order in addition to −1 otherwise.

E. The Apriori Algorithm

One of the most popular data mining approaches is to find frequent item sets from a transaction dataset and derive association rules. Finding frequent item sets (item sets with frequency larger than or equal to a user specified minimum sup-port) are not trivial because of its combinatorial explosion. Once frequent item sets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Apriori is a seminal algorithm for finding frequent item
sets using candidate generation. It is characterized as a level-wise complete search algorithm using antimonotonicity of item sets, “if an item set is not frequent, any of its superset is never frequent”. By convention, Apriori assumes that items within a transaction or item sets are sorted in lexicographic order. Let the set of frequent item sets of size \( k \) be \( F_k \) and their candidates be \( C_k \). Apriori first scans the database and searches for frequent item sets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement.

IV. FUTURE ALGORITHMS

A. Decision tree

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called “root” that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more subspaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute’s value. In the case of numeric attributes, the condition refers to a range. Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path.

B. Naive Bayes Algorithm

Naive Bayes Classifier is a term dealing with simple probabilistic classifier based on applying Bayes Theorem with strong independence assumptions. It assumes that the presence or absence of particular feature of a class is unrelated to the presence or absence of any other feature. The Naive Bayes algorithm is based on conditional probabilities.

It uses Bayes' theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes' theorem can be stated as follows.

\[
\text{Probability (B given A)} = \frac{\text{Probability (A and B)}}{\text{Probability (A)}}
\]

To calculate the probability of B given A, the algorithm counts the number of cases where A and B occur together and divides it by the number of cases where A occurs alone.

C. The Apriori Algorithm

One of the most popular data mining approaches is to find frequent item sets from a transaction dataset and derive association rules. Finding frequent item sets (item sets with frequency larger than or equal to a user specified minimum support) are not trivial because of its combinatorial explosion. Once frequent item sets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Apriori is a seminal algorithm for finding frequent item sets using candidate generation. It is characterized as a level-wise complete search algorithm using anti monotonicity of item sets, “if an item set is not frequent, any of its
superset is never frequent”. By convention, Apriori assumes that items within a transaction or item sets are sorted in lexicographic order. Let the set of frequent item sets of size $k$ be $F_k$ and their candidates be $C_k$. Apriori first scans the database and searches for frequent item sets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement.

V. EXPERIMENT AND RESULT

<table>
<thead>
<tr>
<th>Algorithm Used</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>83 above</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>74 above</td>
</tr>
<tr>
<td>Apriori</td>
<td>73 above</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

There are different data mining classification techniques can be used for the identification and prevention of diabetes and cholesterol diseases among patients. This paper describes some classification techniques in data mining to predict those two diseases in patients. Namely: Decision Tree, Naïve Bayes, and Apriori. These techniques are compared by disease among patients using three classification algorithms accuracy. Then one has highest accuracy above 83%. The Decision Tree algorithm best among three. They are used in various healthcare units all over the world. In future to improve the performance of these classification.

REFERENCES