INTERNATIONAL JOURNAL OF
SCIENCE TECHNOLOGY AND HUMANITIES

Full Length Article

Ascorbic acid in dermatological applications

M.Saktheeswari [a,*], M.Neelamegan [a], M.Nandhini [a]

[a] *Department of Computer Science and Applications, Sri Vidya Mandir Arts and Science College, Katteri, Uthangarai, Tamilnadu, India.*

***Corresponding Author**
ncssaranya@gmail.com
(M.Saktheeswari)
Tel.: +91 9597381128

Received : 19-7-2017
Reviewed: 25-7-2017
Revised : 26-7-2017
Accepted : 05-8-2017

DOI:
https://doi.org/10.26524/ijsth1824

Abstract: Databases are rich with unknown information that can be used for intelligent decision making. Classification and prediction are two form of data analysis that can be used to extract models describing important data classes or to predict future trends. Such analysis can help provide us with a better understanding of the data at a large. Whereas classification predicts definite labels, prediction models continuous valued functions. Several major kinds of classification algorithms including C4.5, ID3, k-nearest neighbor, Naive Bayes and SVM are used for classification. This paper provides a Comparative study of different classification algorithms and their advantages and disadvantages.

Keywords: C4.5, ID3, K-Nearest Neighbor Classifier, SVM, ANN, Naive Bayes.

Introduction

Data Mining is the process of discovering hidden or unknown patterns in huge datasets that are potentially useful and ultimately understandable. The goal of data mining is to extract useful information from huge data sets and to store it as an understandable and structured model for future use, using combined technique of statistics, machine learning and database systems. Data Mining algorithms learns from training datasets to build the model, that can be used on unknown data for prediction. Here we will discuss classification technique that is often called as supervised learning technique .Classification algorithm is used to predict categorical class label of a given data instance so as to classify it into a predetermined class. It is a two step process, in first step classification algorithm uses training data to build a classifier and then in second step it uses this classifier to predict the class label of unlabeled data instance. The classifier is like a function that maps a data instance to a label. The classifier that we will discuss here are C4.5, ID3, k-nearest neighbor, Naive Bayes, ANN and SVM.

Classification and Prediction have numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis.

2. C4.5 ALGORITHM

The C4.5 algorithm improves ID3 in the following ways:

Missing data: When the decision tree is built, missing data are simply unobserved. That is the gain ratio is calculated by the other records that have a value for that attribute.

Continuous data: The basic idea is to divide the data into ranges based on the attribute values for that item that are found in the training sample. Pruning: There are two primary pruning strategies proposed in C4.5 1) Subtree replacement, a subtree is replaced by a leaf node if this replacement results in an error rate close to that of the original tree.

2) Subtree raising, replaces a subtree by its most used subtree. Here a subtree is raised from its current location to a node higher up in the tree.

Rules: C4.5 allows classification via either decision tree or rules generated from them.

Spliting: This approach uses the GainRatio as opposed to gain.The GainRatio is defined as

Gain Ratio (D,S) = Gain(D,S)/ H(|D_1|/|D|,.....|D_s|/|D|)

For splitting purposes, C4.5 uses the largest Gain Ratio that ensures a larger than average information gain.

3. ID3 Algorithm

The basic strategy used by ID3 is to choose splitting attributes with the highest information gain first. The quantity of information associated with an attribute value is related to the probability of event. The concept used to quantify information is called entropy. Entropy is used to measure the amount of uncertainty or chance in a set of data. When all data in a set belong to a single class, there is no uncertainty. In this case the entropy is zero(0).The objective of decision tree classification is to iteratively partition the given data set into subsets belong to the same class.

Definition: Given probabilities $p_1,p_2,.....p_s$ where $\sum \sum s_{i=1} p_i = 1$, entropy is defined as

$$H(p_1,p_2,.....p_s) = \sum s_{i=1} (p_i \log(1/p_i))$$

4. K- Nearest Neighbor Classification (Knn)

The K-Nearest Neighbor Algorithm is the simplest of all machine learning algorithms. It is based on the principle that the samples that are similar, generally lies in close vicinity [1].K-Nearest Neighbor is instance based learning method. Instance based classifiers are also called lazy learners as they store all of the training samples and do not build a classifier until a new, unlabeled sample needs to be classified [2]. Lazy-learning algorithms require less computation time during the training phase than eager learning algorithms (such as decision trees, neural networks and bayes networks) but more computation time during the classification process [3-4]. Nearest-neighbor classifiers are based on learning by resemblance, i.e. by comparing a given test sample with the available training samples which are similar to it. For a data sample X to be classified, its K-nearest neighbors are searched and then X is assigned to class label to which majority of its neighbors belongs to. The choice of k also affects the performance of k-nearest neighbor algorithm [5]. If the value of k is too small, then K-NN classifier may be vulnerable to over fitting because of noise present in the training dataset. On the other hand, if k is too large, the nearest-neighbor classifier may misclassify the test sample because its list of nearest neighbors may contain some data points that are located far away from its neighborhood. K-NN fundamentally works on the belief that the data is connected in a feature space. Hence, all the points are considered in order, to find out the distance among the data points. Euclidian distance or Hamming distance is used according to the data type of data classes used [6]. In this a single value of K is given which is used to find the total number of nearest neighbors that determine the class label for unknown sample. If the value of K=1, then it is called as nearest neighbor classification.

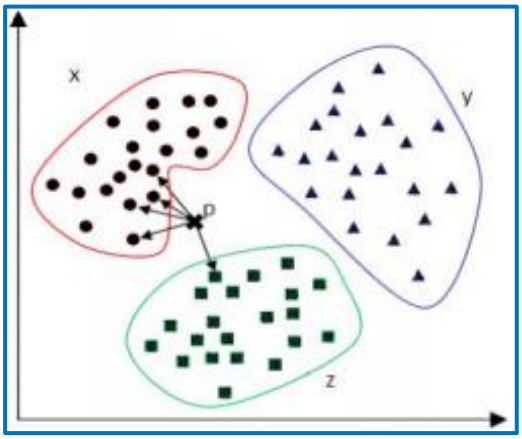

Figure 1: An example of K-NN classifier [7]

The K-NN classifier works as follows:

1. Initialize value of K.
2. Calculate distance between input sample and training samples.
3. Sort the distances.
4. Take top K- nearest neighbors.
5. Apply simple majority.
6. Predict class label with more neighbors for input sample.

Following example shows that there are three classes X, Y and Z as shown in figure 1. Now, it is required to find out the class label for data sample P. Here, value of K=5 and the Euclidean distance is calculated for each sample pair and it is found that four nearest neighbor samples are falling in the class label X, while single tuple belongs to class label Z. So, the sample P is assigned to class X as it is the principal class for that sample.

Advantages

- Easy to understand and implement.
- Training is very fast.
- It is robust to noisy training data.
- It performs well on applications in which a sample can have many class labels [5].

Disadvantages

- Lazy learners incur expensive computational costs when the number of potential neighbors which to compare a given unlabeled sample is large [5].
- It is sensitive to the local structure of the data [8].
- Memory limitation
- As it is supervised lazy learner, it runs slowly.

5. Naive Bayes Classification

Naive Bayes Classifier is the simple Statistical Bayesian Classifier [9]. It is called Naive as it assumes that all variables contribute towards classification and are mutually correlated. This assumption is called class conditional independence [10]. It is also called Idiot's Bayes, Simple Bayes, and Independence Bayes. They can predict class membership probabilities, such as the probability that a given data item belongs to a particular class label. A Naive Bayes classifier considers that the presence (or absence) of a particular feature (attribute) of a class is unrelated to the presence (or absence) of any other feature when the class variable is given. The Naive Bayes Classifier technique is based on Bayesian Theorem and it is used when the dimensionality of the inputs is high [11]. Bayesian classification is based on Bayes Theorem and Bayes Theorem is stated as below: Let

X is a data sample whose class label is not known and let H be some hypothesis, such that the data sample X may belong to a specified class C. Bayes theorem is used for calculating the posterior probability P (C|X), from P(C), P(X) and P(X|C). Where

P (C|X) is the posterior probability of target class. P(C) is called the prior probability of class.

P (X|C) is the likelihood which is the probability of predictor of given class.

P(X) is the prior probability of predictor of class.

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)}$$

The Naive Bayes classifier [7] works as follows:

1. Let D be the training dataset associated with class labels. Each tuple is represented by n-dimensional element vector, X=(x1, x2, x3,.....,xn).
2. Consider that there are m classes C1, C2, C3...., Cm. Suppose that we want to classify an unknown tuple X, then the classifier will predict that X belongs to the class with higher posterior probability, conditioned on X. i.e., the Naive Bayesian classifier assigns an unknown tuple X to the class Ci if and only if P(Ci|X) > P(Cj|X) For $1 \le j \le m$, and $i \ne j$, above posterior probabilities are computed using Bayes Theorem.

Advantages

- It requires short computational time for training.
- It improves the classification performance by removing the irrelevant features.
- It has good performance.

Disadvantages

- The Naive Bayes classifier requires a very large number of records to obtain good results.
- Less accurate as compared to other classifiers on some datasets.

6. ANN (Artificial Neural Network) Algorithm

Neural networks are often referred to as artificial neural networks (ANN). The NN is an information processing system that consist of a graph representing the processing system and various algorithm that access that graph. The NN can be viewed as a directed graph with source (output), (input), s ink (output) and internal (hidden) nodes. To perform the data mining task, a tuple is input through the input nodes and the output node determines what the prediction is. Unlike decision tree, which has only one input node, the NN has one input node for each attribute value to be examined to solve the data mining function. A neural network (NN) model is a computational model consisting of three parts:

1. Neural network graph that defines the data structure of the neural network.
2. Learning algorithm that indicates how learning takes place.

Recall techniques that determine how information is obtained from the network. NNs have been used in speech recognition and synthesis, pattern recognition, medical applications, fault detection, problem diagnosis robot control and computer vision.

7. SVM (Support Vector Machine) Algorithm

Support Vector Machine (SVM) is based on statistical learning theory and is increasingly becoming useful in data mining. The main idea is to non-linearly map the dataset into a high-dimensional feature space and use a linear discriminator to classify the data. Its success has been demonstrated in the areas of regression, classification and decision tree construction.

Unlike other classification techniques, which attempt to minimize error of classification, SVMs incorporate structured risk minimization which minimizes an upper bound on the generalized error. Consider two sets A and B are linearly separable. The idea is to determine from an infinite number of planes correctly separating A and B, the one which will have the smallest generalization error. SVMs select with which minimizes the margin separating the two classes. The margin is defined as the distance between the separating hyperplane to the nearest point of a, plus the distance from the hyperplane to the nearest point in B.

8. Conclusion

This paper deals with different classification techniques used in data mining and a study of them. These classification algorithms can be implemented on variety of data sets like patient data set, share market data set etc. Hence these classification techniques show how data can be grouped and determined when a new set of data is available. Each technique has both advantages and disadvantages which are given in the paper. Based on the conditions each one as needed can be selected.

References

[1]. T. M. Cover and P. E. Hart, Nearest Neighbor Pattern Classification, *IEEE Transactions on Information Theory*, 13 (1967) 21-27.

[2]. J. Han, M. Kamber, Data Mining Concepts and Techniques, *Elevier*, (2011) 744.

[3]. K. P. Soman, (2006) Insight into Data Mining Theory and Practice, *Prentice Hall India Learning Private Limited*, New Delhi: PHI.

[4]. S. B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, *Informatica*, 31 (2007) 249-268.

[5]. H. Bhavsar, A. Ganatra, A Comparative Study of Training Algorithms for Supervised Machine Learning, *International Journal of Soft Computing and Engineering (IJSCE)*, 2 (2012) 74-81.

[6]. D. Michie, D.J. Spiegelhalter, C.C. Taylor (1994) Machine Learning, Neural and Statistical Classification, Ellis Horwood Upper Saddle River, NJ, USA.

[7]. Bhavesh Patankar, Dr. Vijay Chavda, A Comparative Study of Decision Tree, Naive Bayesian and k-nn Classifiers in Data Mining, *International Journal of Advanced Research in Computer Science and Software Engineering*, 4 (2014) 776-779.

[8]. S.Archana, Dr. K.Elangovan, Survey of Classification Techniques in Data Mining, *International Journal of Computer Science and Mobile Applications*, 2 (2014) 65-71.

[9]. R. Duda, P. Hart, (1973) Pattern Classification and Scene Analysis, *John Wiley and Sons*, New York.

[10]. N. Friedman, D. Geiger, Goldazmidt, Bayesian Network Classifiers, *Machine Learning*, 29 (1997) 131-163.

[11]. Sagar S. Nikam, A Comparative Study of Classification Techniques in Data Mining Algorithms, *Oriental Journal of Computer Science and Technology*, 8 (2015) 13-19.

[12]. T.Joachims, Making large-scale support vector machine learning practical, *In Advances in Kernel Methods: Support Vector Machines*, (1999) 169-184.

[13]. Delveen Luqman Abd Al.Nabi, Shereen Shukri Ahmed, Survey on Classification Algorithms for Data Mining: (Comparison and Evaluation), *Computer Engineering and Intelligent Systems*, 4 (2013) 18-24.

[14]. Arun K.Pujari, (2013) Data Mining Techniques, *Universities Press (India)Pvt. Ltd*, USA.

[15]. Margaret H. Dunhan, (2013) Data Mining: Introductory and Advanced Topics, *Pearson Education*.