Original Research Paper

VOLUME-6 | ISSUE-3 | MARCH - 2017 • ISSN No 2277 - 8179 | IF : 4.176 | IC Value : 78.46

# A Comparative Study on Classification Using Hypothyroidism

Statistic

**KEYWORDS:** Classification, Hypothyroidism KNN Classification,

**Gothai Nachiyar S** — Research Scholar Manonmaniam Sundaranar University ,Tirunelveli.

**K. Senthamarai Kannan** — Professor & Head ,Manonmaniam Sundaranar University

## ABSTRACT

Classification is one of the important role in the field of healthcare. Diagnosis of health conditions is a very important and challenging task in field of medical science. There are various types of diseases are diagnosis in medical science. Thyroid disease is one of critical diseases that is very serious problem and affected the health of human being. Thyroid diseases classification is one of the important problems in medical science because it is directly related to health condition of human body, this type of disease can be solve by proper identify and carefully treatment. There are various authors have worked in the field of thyroid diseases classification and give the classification accuracy with different model. In this work, KNN classification, Discriminant Analysis and Vector support machine are applied for classification of thyroid data. This research is focus on comparing the classification technique with regards to the classification accuracy.

## INTRODUCTION

Data mining is a young and promising field of information and knowledge discovery (Han et al., 2011). Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. Academicians are using data-mining approaches like decision trees, clusters, neural networks, and time series to publish research. Healthcare, however, has always been slow to incorporate the latest research into everyday practice. Deneshkumar et al., (2014) have discussed the large quantities of information about patients and their medical conditions through data mining techniques. Chein and Chen (2006) used several attributes to predict the employee performance. They specified age, gender, marital status, experience, education, major subjects and school tires as potential factors that might affect the performance. Then they excluded age, gender and marital status, so that no discrimination would exist in the process of personal selection. As a result for their study, they found that employee performance is highly affected by education degree, the school tire, and the job experience. Kahya (2007) also searched on certain factors that affect the job performance. The researcher reviewed previous studies, describing the effect of experience, salary, education, working conditions and job satisfaction on the performance. As a result of the research, it has been found that several factors affected the employee's performance.

Data mining has become a fundamental methodology for computing applications in medical informatics. Progress in data mining applications and its implications are manifested in the areas of information management in healthcare organizations, health informatics, epidemiology, patient care and monitoring systems, assistive technology, large-scale image analysis to information extraction and automatic identification of unknown classes. **Classification** is a **data mining** function that assigns items in a collection to target categories or classes. The goal of **classification** is to accurately predict the target class for each case in the **data**. Milan and Godara (2011) compared data mining classification techniques RIPPER classifier, Decision Tree, Artificial neural networks (ANNs), and Support Vector Machine (SVM) are analyzed on cardiovascular disease dataset. In this paper, KNN, SVM and discriminant analysis were classified and were compared for classification accuracy.

## 2. MATERIALS AND METHODS

We collected a multi-dimensional healthcare dataset. This multidimensional thyroid disease dataset contains 45 samples with 3 variables (Symptom score, T4 and TSH). This dataset is collected from the journal Khanale and Ambilwade (2011). All 45 observations were classified subclinical, secondary, primary and nohypo using KNN, SVM and Discriminant Analysis.

### 2.1 K-NEAREST NEIGHBOR

K-nearest neighbor is a supervised learning algorithm developed by Hodges in 1967, where the result of new instance query is classified based on majority of K-nearest neighbor category. The purpose of this algorithm is to classify a new object based on memory. Given a query point, we find k number of objects or (training points) closest to the query point. The classification is using majority vote among the classification of the k-objects. Any ties can be broken at random. K nearest neighbor algorithm used neighborhood classification as the prediction value of the new query instance. K nearest neighbor algorithm is very simple. It works based on minimum distance from the query instance to the training samples to determine the k nearest neighbors David et al. (2008).

### K-NEAREST NEIGHBOR ALGORITHM

The main steps of the KNN algorithm are as follows:

1. Assume there are N training objects where each object has t-attributes and an object can belong to one of the m-classes.

2. Let O be an object to be classified.

3. Compute the distance between the object O and each of the training objects.

4. Let $d_1, d2_{....} d_N$ be the resulting distances.

5. Arrange the distances in ascending order and identify first k objects corresponding to the first k smallest distances to get the set $C_k$.

6. Let $x_r$ denote the number of objects in the set $C_k$ belong to the class $r^* = 1,2,...,m)$. We assign the object to the class $\lambda$ if, $\pi\lambda_= max_{\{x_1, x_{2....}, x_m\}}$

### 2.2 SUPPORT VECTOR MACHINE

The support vector machine (SVM) is a training algorithm for learning classification and regression rules from data, SVMs were first suggested by Vapnik in the 1960s for classification and have recently become an area of intense research owing to developments in the techniques and theory coupled with extensions to regression and density estimation. SVMs arose from statistical learning theory; the aim being to solve only the problem of interest without solving a more difficult problem as an intermediate step. SVMs are based on the structural risk minimisation principle, closely related to regularisation theory.

### SVM ALGORITHM

The main steps of the SVM algorithm are as follows:

1. Let X = (X1, . . . ,Xp) 2 X = Rp denote the predictors

2. Y the variable for the categorical outcome which takes one of, say, k

nominal class labels, Y = {1,. . . , k}. Then the so-called training data are given as a set of n observation pairs,

3. Dn = {(xi, yi), i = 1, . . . , n}, where(xi, yi)'s are viewed as independent and identically distributed random outcomes of (X, Y )from some unknown distribution PX,Y .
4. The ultimate goal of classification is to understand informative patterns that exist in thepredictors in relation to their corresponding class labels. For that reason, classification is known as pattern recognition in the computer science literature.

5. Formally, it aims to find amap (classification rule), _ : X ! Y based on the training data which can be generalized to future cases from the same distribution PX,Y .

6. For simplicity, this section is focused on classification with binary outcomes only (k = 2).Typically, construction of such a rule _ is done by finding a real-valued discriminant function.

7. f first and taking either the indicator _(x) = I( f(x) _ 0) if two classes are labeled as 0 or1, or its sign _(x) = sgn( f(x))if they are symmetrically labeled as}1. In the latter, the classification boundary is determined by the zero level set of f, i.e. {x : f(x) = 0}.

**2.3 DISCRIMINANT ANALYSIS**

Discriminant Analysis is a statistical tool with an objective to assess the adequacy of a classification, given the group memberships; or to assign objects to one group among a number of groups. For any kind of Discriminant Analysis, some group assignments should be known beforehand.Discriminant Analysis is quite close to being a graphical version of MANOVA and often used to complement the findings of Cluster Analysis and Principal Components Analysis.

When Discriminant Analysis is used to separate two groups, it is called Discriminant Function Analysis (DFA); while when there are more than two groups – the Canonical Varieties Analysis (CVA) method is used.In the 1930's, 3 different people – R.A. Fisher in UK, Hoteling in US and Mahalanob is in India were trying to solve the same problem via three different approaches. Later their methods of Fisher linear discriminant function, Hoteling's T2 test and Mahalanobis D2 distance were combined to devise what is today called Discriminant Analysis.

**DISCRIMINANT ANALYSIS ALGORITHM**

- o Step 1: Computing the d-dimensional mean vectors
- o Step 2: Computing the Scatter Matrices
  - 2.1 Within-class scatter matrix SWSB
  - 2.2 Between-class scatter matrix SWSB
- o Step 3: Solving the generalized eigenvalue problem for the matrix SW−1SB
  - Checking the eigenvector-eigenvalue calculation
- o Step 4: Selecting linear discriminants for the new feature subspace
  - 4.1. Sorting the eigenvectors by decreasing eigenvalues
  - 4.2. Choosing k eigenvectors with the largest eigenvalues
- o Step 5: Transforming the samples onto the new subspace

Below formulae were used to calculate accuracy:
Accuracy = (TP + TN) / (TP + FP + TN + FN)

**3. Computational Results and Discussions**

**Table 3.1: Classification Using KNN**

| KNN | subclinical | Secondary | Primary | nohypo | Total |
|---|---|---|---|---|---|
| Subclinical | 14 | 0 | 1 | 1 | 16 |
| Secondary | 0 | 10 | 2 | 1 | 13 |
| Primary | 1 | 1 | 4 | 1 | 7 |
| Nohypo | 1 | 1 | 1 | 6 | 9 |

**Table 3.2: Classification Using SVM**

| SVM | Subclinical | Secondary | Primary | nohypo | Total |
|---|---|---|---|---|---|
| Subclinical | 15 | 0 | 1 | 0 | 16 |
| Secondary | 1 | 11 | 1 | 0 | 13 |
| Primary | 0 | 2 | 5 | 0 | 7 |
| Nohypo | 2 | 0 | 1 | 6 | 9 |

**Table 3.3: Classification Using Discriminant Analysis**

| Discriminant Analysis | Subclinical | secondary | Primary | Nohypo | Total |
|---|---|---|---|---|---|
| Subclinical | 10 | 1 | 1 | 4 | 16 |
| Secondary | 2 | 9 | 1 | 1 | 13 |
| Primary | 1 | 1 | 5 | 0 | 7 |
| Nohypo | 0 | 1 | 2 | 6 | 9 |

The classification are as follows:

- Out of sixteen observations using KNN, 14 were correctly classified from subclinical to subclinical, and others were misclassified. Similarly using SVM 15 were correctly classified and other were misclassified and SVM, 10 were correctly classified from subclinical to subclinical and others were misclassified.
- Out of thirteen observations using KNN ten were correctly classified other were misclassified. SVM using eleven observations were correctly classified from 'secondary to secondary and other were misclassified. Discriminant analysis nine were correctly classified and other were misclassified.
- Out of seven observations, using KNN four observations were correctly classified from primary to primary' and others were misclassified. Using SVM, five observations were correctly classified other were misclassified. Using Discriminant Analysis five were correctly classified and other were misclassified.
- Out of nine observations using KNN, six observations were correctly classified nohypo to nohypo' and other were misclassified. Using SVM, six observations were correctly classified and others were misclassified. Using Discriminant analysis, six observations were correctly classified and others were misclassified.

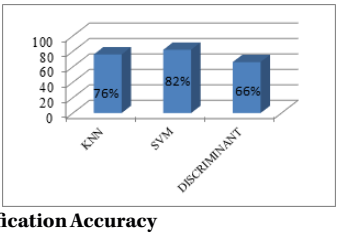


**Fig. 1: Classification Accuracy**

**4. Conclusion**

This work focused the implementation of k-Nearest Neighbor, Discriminant Analysis and support vector machines for the hypothyroidism data. From the analysis, it is examined that the formation of classifications will be different for three classification methods. From the Barchart, it is seen that the Support Vector Machine accuracy is 82%, KNN gives the accuracy of 76% Discriminant analysis gives the accuracy of 66 %. Support vector machines are higher than the accuracy of KNN and Discriminant analysis.

**REFERENCES:**

1. Chein, C., Chen, L. (2006) "Data mining to improve personnel selection and enhance human capital: A case study in high technology industry", Expert Systems with Applications, In Press.
2. Deneshkumar V., Senthamarai Kannan K., and Manikandan M. (2014): Identification of Outliers in Medical Diagnosis System Using Data Mining Techniques. International Journal of Statistics and Applications, Vol.4, No.6, pp.241-248.
3. Han, J., Kamber, M., Jian P. (2011). Data Mining Concepts and Techniques. San Francisco, CA: Morgan Kaufmann Publishers.
4. Kayha, E. (2007). "The Effects of Job Characteristics and Working Conditions on Job Performance", International Journal of Industrial Ergonomics, In Press.
5. Khanale .P.B and Ambilwade R.P (2011). " A fuzzy Inference System for Diagnosis of Hypothyroidism". Journal of Artificial Intelligence 4(1): 45 -54.
6. Milan Kumari, Sunila Godara(2011). "Comparative Study of Data Mining Classification Methodsin Cardiovascular Disease Prediction". IJCST Vol. 2, Issue 2