

Ditapol Muntham 2010: Applications of Decision Tree Algorithms with Diagnosis of the Respiratory System. Master of Science (Statistics), Major Field: Statistics, Department of Statistics. Thesis Advisor: Assistant Professor Lily Ingsrisawang, Ph.D. 126 pages.

The objectives of this study involved: 1) the application of methods of knowledge discovery from data base using decision tree algorithms for Respiratory System Diagnosis to classify patient at Pranakornsri Ayudthaya Hospital into three groups: acute URI, acute sinusitis, and pneumonia, and 2) the comparison of the performance of three decision tree algorithms : ID3, C4.5, and CART, for the classification or screening the patients with the three diseases. The data used in this study came from the medical records of 7,327 out-patients with respiratory diseases at Pranakornsri Ayudthaya Hospital in the period of 2003 to 2006. The variables considered were age, body temperature, residence, occupation, and some symptoms such as rhinorrhea, fever, nasal congestion, periorbital pain, headache, wheezing and coughing. The study methods were the knowledge discovery with the employment of ID3, C4.5, and CART decision tree algorithms from hospital's patient registertion database and the effective comparison of the three algorithms. The algorithms used for the important variable selection were Best First, Greedy Stepwise, and Genetic Search. The decision tree algorithms were employed to classify the patients with those three diseases into three groups. The validity of the decision tree algorithms was studied by the division of the data into two sets, a training and a testing data sets, which were based on the cross-validation and the percentage split methods.

The study of the methods of knowledge discovery for the classification found that, for the patients with acute URI with only 7 selected variables and the ratio 70:30 of the training data set and the testing data set, the C4.5 algorithm was the most effective. The classification accuracy was 92.31%; the sensitivity was 90.60%; the PPV was 95.63%; the NPV was 90.18%; and the AUC was 96.20%. For the classification of the patients with acute sinusitis with only 8 selected variables and the ratio 70:30 of the training data set and the testing data set, the C4.5 algorithm was the most effective. The classification accuracy was 94.70%; the sensitivity was 93.42%; the PPV was 83.24; the NPV was 98.16%; and the AUC was 92.20%. For the classification of the patients with pneumonia with only 7 selected variables and the ratio 50:50 of the training data set and the testing data set, the CART algorithm was the most effective. The classification accuracy was 94.69%; the sensitivity was 100%; the PPV was 81.57; the NPV was 100%; and the AUC was 94.70%. The obtained results could be used for the support of the diagnosis of patients with Respiratory System disease.

---

Student's signature

---

Thesis Advisor's signature