





Prognosis of Thyroid Disease Using MS-Apriori Improved Decision Tree

Yuwei Hao^{1,2} , Wanli Zuo^{1,2}, Zhenkun Shi^{1,2} , Lin Yue³,
Shuai Xue⁴, and Fengling He^{1,2}

¹ College of Computer Science and Technology, Jilin University,
Changchun 130012, China

shizk14@mails.jlu.edu.cn

² Key Laboratory of Symbol Computation and Knowledge Engineering,
Jilin University, Ministry of Education, Changchun 130012, China

³ School of Computer Science and Information Technology,
Northeast Normal University, Changchun 130117, China

⁴ The First Hospital of Jilin University, Changchun 130021, China

Abstract. The lymph nodes metastasis in the papillary thyroid microcarcinoma (PTMC) can lead to a recurrence of cancer. We hope to take preventive measures to reduce the recurrence rate of the thyroid cancer. This paper presents a decision tree improved by MS-Apriori for the prognosis of lymph node metastasis (LNM) in patients with PTMC, called MsaDtd (Decision tree Diagnosis based on MS-Apriori). The method converts the original feature space into a more abundant feature space, MS-Apriori is used to generate association rules that consider rare items by multiple supports and fuzzy logic is introduced to map attribute values to different subintervals. Then, we filter the ranked rules which consider positive and negative tuples. We improve accuracy through deleting disturbance rules. At last, we use the decision tree to predict LNM by analyzing the affiliation between the instance and rules. Clinical-pathological data were obtained from the First Hospital of Jilin University. The results show that the proposed MsaDtd achieves better prediction performance than other methods on the prognosis of LNM.

Keywords: MS-Apriori · Decision tree · Medical mining · Disease predication

1 Introduction

Artificial intelligence (AI) has recently gained a tremendous advance in various applications, e.g., autonomous driving, big data, pattern recognition, intelligent search, image understanding, automatic programming, and robotics [1]. These applications also inspire AI technique to develop and innovate, in a way. The increasing availability of healthcare data and rapid development of big data analytic methods have made possible the recent successful applications of AI in healthcare [2]. Machine Learning as one of the core technologies of AI has been widely used in all walks of life. In recent years, the healthcare industry produces a huge amount of digital data by utilizing information from all sources of healthcare data such as Electronic Health Records [3] and Personal Health

Records [4]. At the same time, machine learning is well poised to assist clinical researchers in deciphering complex predictive patterns in healthcare data [5]. All of these provides the basis of the prognostication of diseases with Machine Learning technique.

Indeed, the incidence of thyroid cancer has nearly tripled since 1975 [6]. In PTMC, the prevalence of subclinical CLNM has been detected as 30%–65% [7]. And PTMC can lead to a recurrence of cancer. Therefore, it is urgent to introduce machine learning into the field of Thyroid Disease. To solve the prognostic problem of Thyroid Disease, we propose a disease diagnosis model, and apply it to thyroid disease diagnosis in the First Hospital of Jilin University.

The technical contributions done in this paper are summarized as follows:

1. We propose an algorithm MsaDtd that converts the original characteristic space into a larger characteristic space and improved decision tree algorithm for disease diagnosis to predict LNM in patients with PTMC.
2. We use MS-Apriori to obtain composite features, taking into account rare items by setting multiple minimum supports (MIS), and introduce fuzzy logic to deal with continuous attributes, aiming to avoid the cost of producing large frequent items.
3. We use 5425 Clinical-pathological data of PTMC patients in the First Hospital of Jilin University to validate MsaDtd. Experimental analysis indicates that the algorithm predicts LNM effectively and accurately.

2 Related Work

Prediction of thyroid diseases using machine learning has been an ongoing effort in recent years. Chen et al. [8] presented a three-stage expert system based on a hybrid support vector machines (SVM). It combined feature selection and parameter optimization, the developed FS-PSO-SVM expert system achieved excellent performance in distinguishing among hyperthyroidism, hypothyroidism and normal ones. Makas et al. [9] developed seven distinct sorts of Neural Networks to identify the thyroid disease. And used particle swarm optimization (PSO), artificial bee colony (ABC) and migrating birds optimization (MBO) algorithms retrained the network. The accuracy of the network developed outperformed the similar studies. Pourahmad et al. [10] used a back propagation feedforward neural networks to diagnose the malignancy in thyroid tumor. Thirteen batch learning algorithms were investigated and three different numbers of neuron in hidden layer were compared to achieve the best performance. Kaya et al. [11] applied Extreme Learning Machine (ELM) to the diagnosis of thyroid disease. This study indicated the classification and speed of ELM were higher than other machine learning methods. Maysanjaya et al. [12] used Multilayer Perceptron method to identify the type of thyroid (normal, hypothyroid, hyperthyroid) with WEKA tool. The accuracy of the prediction was as high as 96.74%.

Researchers have done much research on solving the problem of thyroid diseases diagnosis. But there are few studies on the prognosis of LNM in patients with PTMC. The prognosis of LNM is essential to prevent recurrence of cancer. For the above situation, this paper designs an intelligent decision model MsaDtd to predicts lymph node metastasis (LNM) in patients with PTMC.

3 The Prognosis Algorithm Based on MS-Apriori and Decision Tree

We design a disease diagnostic algorithm by mapping the prognosis of LNM in patients with PTMC to a binary classification problem. The symptoms of patients are mapped to independent variables $\mathbf{u} = (u_1; u_2; \dots; u_d)$ and diagnostic results are mapped to dependent variables $y \in \{0, 1\}$.

3.1 MS-Apriori Rule Mining

Apriori play a major role in identifying frequent itemset and deriving rule set out of it [13]. Using Apriori results in a shortage when mining rare knowledge patterns of rare events, due to the entire database only set one minimum support. To solve this problem, we use MS-Apriori setting MIS for different items.

For attribute value, this paper introduces fuzzy logic to map attribute values to different subintervals through membership function, aiming to avoid the cost of producing large frequent items.

The association rule mining process is as follow. An item type v_i is defined as each value type under each attribute in clinical-pathological data. The set of items in the whole database is I shown in Eq. (1) and the item type set is V shown in Eq. (2).

$$I = \{a_1, a_2, \dots, a_m\} = IA_1 \cup IA_2 \cup \dots \cup IA_d, m = n * d \tag{1}$$

$$V = \{v_i\}, i = 1, 2, \dots, u \tag{2}$$

There are m items, u item types and d -dimension features in the whole database. $IA = \{a_i\}$ ($i = 1, 2, \dots, n$) represents the fuzzy itemset under an attribute. We specify that different attribute values under the same attribute do not belong to the same frequent itemset. The items in a frequent itemset should meet the condition shown in Eq. (3).

$$a_x \cap a_y = \emptyset, a_x \in IA_i, a_y \in IA_j, i = j \tag{3}$$

In addition, the support of frequent itemset is the smallest MIS of items in the frequent itemset. The frequent itemset is defined as Eq. (4). The MIS of frequent itemset c is defined as Eq. (5). The MIS of the item is defined as Eq. (6).

$$c = \{a_1, a_2, \dots, a_k\}, 1 \leq k \leq d \tag{4}$$

$$MIS(c) = \min(MIS(a_1), MIS(a_2) \dots MIS(a_k)) \tag{5}$$

$$MIS(v_i) = \frac{v_i \cup LM_{yes}}{N} \tag{6}$$

v_i represents an item, corresponding a value type in clinical-pathological data. LM_{yes} represents the label of patients is lymph node metastasis. N is the total number of instances. The probability of item v_i and item LM_{yes} appear in the same frequent itemset is set to the MIS of v_i .

The frequent item c_j is converted to rule $Rule_j$ shown in Eqs. (7) and (8).

$$c_j : a_1 \cup a_2 \cup \dots \cup LM_{yes}/LM_{no} \quad (7)$$

$$Rule_j \rightarrow LM_{yes}, Rule_j : a_1 \cup a_2 \cup \dots \cup a_{k-1} \quad (8)$$

We rank the rule by cosine measure and delete disturbance rules by defining a threshold. The cosine measure of positive tuple rules is defined by Eq. (9).

$$\text{cosine}(Rule_j, LM_{yes}) = \frac{P(Rule_j \cup LM_{yes})}{\sqrt{P(Rule_j) * P(LM_{yes})}} \quad (9)$$

$P(Rule_j \cup LM_{yes})$ represents the probability that $Rule_j$ and LM_{yes} belong to the same frequent item. The cosine measure of negative tuple rules is defined as Eq. (10).

$$\text{cosine}(Rule_j, LM_{no}) = \frac{P(Rule_j \cup LM_{no})}{\sqrt{P(Rule_j) * P(LM_{no})}} \quad (10)$$

Algorithm 1 outlines the process of rule mining by MS-Apriori. SDC is used to limit a rare item and a common item appear in the same frequent item. $threshold$ is used to delete disturbance rules.

Algorithm 1:MS-Apriori rule mining algorithm

Input: LNM Dataset $D = \{(u_i, y_i)\}, i = (1, 2, \dots, n), y_i \in \{0, 1\}$ which contains training instances $u = (u_1; u_2; \dots; u_d)$ and their associated diagnostic labels $y \in \{0, 1\}$, membership functions $\{\delta_m(x)\}, m = (1, 2, \dots, d)$, $SDC, threshold$

Output: sorted rule set which removed disturbance rules:

$R = \{rule \mid \text{cosine}(rule) \geq threshold\}$

1. Convert D to T by converting each attribute value in D to a_i by $\delta_m(x)$
 2. Compute multiple minimum supports for each item by $MIS(v_i) = \frac{v_i \cup LM_{yes}}{N}$
 3. Generate frequent 1 itemset $F_1 = \{< c > \mid c \in T, c.count \geq MIS(c)\}$
 4. **for** ($k = 2; F_{k-1} \neq \emptyset; k++$)
 5. $C_k = \text{gen_candidate}(F_{k-1}, SDC)$
 6. **for each** transaction t in T **do**
 7. Storage t in C_k when t is the subset of C_k
 8. **for each** candidate c in C_k **do**
 9. $c.count++$
 10. Generate $F_k = \{c \in C_k \mid c.count \geq MIS(c)\}$
 11. Generate $\{F_k\}$ ($k=1, 2, \dots, k$)
 12. **for each** c in $\{F_k\}$ **do**
 13. Convert c to Diagnosis rules $rule$
 14. ranking $rule$ and delete the rule when $\text{cosine}(rule) < threshold$
-

3.2 Decision Tree Construction

We obtain the sorted rule set $R = \{rule | cosine(rule) \geq threshold\}$ which is closely related to LMN diagnosis, through mining association rules in clinical-pathological data. Next, we build a decision tree which is used to predict LNM.

Through converting each rule in rule set R to the candidate attributes of the decision tree, the algorithm generates attribute set A . To determine which rule is selected as the splitting attribute in the process of classification, information gain is used as a decision criterion. When an instance contains all items needed in $rule_i$, this rule can be applied to this instance. $rule_i$ as a new attribute, its attribute value is LM_{yes}/LM_{no} . If the rule is positive tuple rule, the value of $rule_i$ is LM_{yes} after applying the rule. If the rule is negative tuple rule, the value of $rule_i$ is LM_{no} after applying the rule. Otherwise, the rule cannot be applied, the value is No. The dataset D is converted to $S = \{(x_i, y_i)\}, i = (1, 2, \dots, n), y_i \in \{0, 1\}$. The labels of the dataset are LNM and normal. We mark it as S_1 and S_0 . The information entropy of S is defined as Eq. (11).

$$H(S) = - \sum_{i=1}^2 p_i \log_2 p_i \quad (11)$$

$$p_i = \frac{S_i}{S}, i = 1, 2 \quad (12)$$

Where p_i represents the probability that $x_i \in S$ belongs to a class S_i , and is estimated by Eq. (12). The information gain for attribute $r \in A$ at node N is defined as Eq. (13).

$$Gain(S, r, N) = H(S) - \sum_{j=0}^1 \frac{S_j}{S} H(S_j) \quad (13)$$

The attribute with the maximum information gain is selected as the splitting attribute at node N . The instances are recursively partitioned into smaller subsets through analyzing the affiliation between instances and the rules mined by MS-Apriori. When all the subsets belong to a single class, or there are no instance or attribute can be used to partition, a model used to predict LNM is constructed.

4 Experiments

4.1 Data Pre-processing

This study is conducted in the Thyroid Surgery of the First Hospital of Jilin University. A total of 5425 patients with PTMC who underwent thyroidectomy with neck dissection from 2011 to 2015 are studied. Among the 5254 patients, there are 4855 cases met the criteria, including 323 cases treated lateral neck dissection.

Features used in this study include gender, age, capsule invasion (CI), maximum tumor diameter (MTD), multifocal, Hashimoto thyroiditis (HT), Central lymph node

Table 1. Description of feature

| Features | Gender | Age | CI | MTD | Multifocal | HT | CN | LN |
|----------|-------------|-------|-----|-----|------------|-----|------|------|
| range | Male/female | 12–82 | 0–1 | 0/1 | 0/1 | 0/1 | 0–34 | 0–87 |

number (CN). These features are shown in Table 1. For LLNM, adding two additional features, CLNM and lateral lymph node number (LN).

In this paper, we use the box plot to analyze data. We identify noise data by IRQ and set the value of it as null. Because box plot identifies abnormal values more objective and quartiles have a certain degree of robustness. For the missing values, in order to avoiding the loss of information by deleting. We should speculate missing values based on the majority of the existing data. We use mean/mode imputation (MMI) to deal with missing values. A bias occurs when we use it to train a predictive model, because of the unbalanced data. To solve the problem of skewed data, we use balancing techniques. The techniques we use is on CNLM dataset is KNN-NearMiss-2, a kind of supervised under-sampling techniques based on K-nearest neighbor. For LLNM dataset, SMOTE over-sampling technique is used, due to the small number of instances.

4.2 Results and Discussion

The proposed predictor is applied to the Clinical-pathological data of the First Hospital of Jilin University. To illustrate the performance of MsaDtd, we compare MsaDtd with a range of baseline algorithms, including Decision Tree (DT), Support Vector Machines (SVM), Logistic regression (LR), Bernoulli Bayes (BNB). We use 10-fold cross-validation to valid MsaDtd algorithm on CLNM dataset and LLNM dataset.

Table 2. Performance comparison with baseline algorithms on CLNM dataset

| | Accuracy | Precision | Recall | F ₁ | AUC |
|--------|----------|-----------|--------|----------------|--------|
| MsaDtd | 76.09% | 72.16% | 63.63% | 72.63% | 82.06% |
| DT | 73.62% | 67.57% | 72.44% | 73.94% | 74.13% |
| SVM | 71.03% | 65.86% | 63.22% | 68.54% | 75.34% |
| LR | 70.58% | 64.97% | 67.27% | 69.56% | 75.37% |
| BNB | 59.05% | 55.38% | 62.88% | 60.54% | 62.32% |

Table 3. Performance comparison with baseline algorithms on LLNM dataset

| | Accuracy | Precision | Recall | F ₁ | AUC |
|--------|----------|-----------|--------|----------------|--------|
| MsaDtd | 87.21% | 82.75% | 85.86% | 86.85% | 88.37% |
| DT | 83.70% | 78.54% | 83.95% | 83.76% | 83.20% |
| SVM | 79.19% | 71.72% | 91.02% | 81.40% | 86.08% |
| LR | 78.79% | 72.11% | 84.43% | 79.80% | 87.31% |
| BNB | 75.08% | 68.53% | 82.38% | 76.74% | 82.42% |

Tables 2 and 3 shows the results of various algorithms on CLNM dataset and LLNM dataset, respectively. As we can see, on CLNM dataset, MsaDtd algorithm

achieves the results with Accuracy, Precision, Recall, F_1 and AUC values are 76.09%, 72.16%, 63.63%, 72.63%, and 82.06%. High prediction accuracy of 76.09% is obtained for MsaDtd algorithm. The accuracy of the improved decision tree is higher than the traditional decision tree and other classifiers. The accuracy of improved decision tree MsaDtd increased by 2.47% compared with the traditional decision tree. On LLNM dataset, the average prediction Accuracy, Recall, Precision, F_1 , and AUC of MsaDtd are 87.21%, 82.75%, 85.86%, 86.85% and 88.37%. Our method outperforms the traditional decision tree in all aspects. The Accuracy, Recall, Precision, F_1 , and AUC increased by 3.51%, 4.21%, 1.91%, 3.09% and 5.17% comparing to the decision tree. Our method has the highest Accuracy, Recall, Precision and AUC among the methods we compared.

Figures 1 and 2 shows a plot of the ROC curves derived from MsaDtd and various baseline algorithms on different dataset. One CLNM dataset, it is higher 6.69% than LR which having the highest ROC area among baseline algorithms. On LLNM dataset, the Roc area of MsaDtd is 88.37%, which is higher than all of the methods mentioned. The above results show the superior performance of the prediction we proposed.

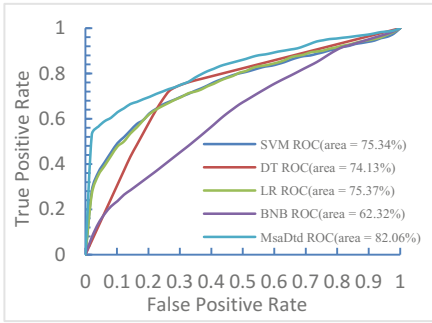


Fig. 1. ROC curve comparison with baseline algorithms on CLNM dataset

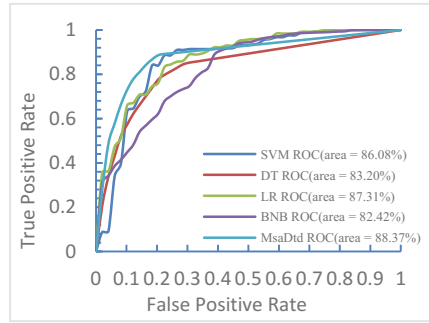


Fig. 2. ROC curve comparison with baseline algorithms on LLNM dataset

Table 4. Performance comparison with DeepPPI on CLNM and LLNM dataset

| Dataset | Method | Accuracy | Precision | Recall | F_1 | AUC |
|---------|---------|----------|-----------|--------|--------|--------|
| CLNM | MsaDtd | 76.09% | 72.16% | 63.63% | 72.63% | 82.19% |
| | DeepPPI | 65.66% | 63.78% | 73.96% | 68.28% | 74.71% |
| LLNM | MsaDtd | 87.21% | 82.75% | 85.86% | 86.85% | 89.15% |
| | DeepPPI | 81.83% | 76.22% | 92.10% | 83.05% | 87.09% |

To our Knowledge, there is almost no one proposed the specialized algorithm for the prognosis of lymph node metastasis (LNM) in patients with PTMC in recent years, so we compare our method with a classification algorithm DeepPPI-Con [14] which achieves superior performance in Protein-Protein Interactions. The results shown in Table 4 indicate that our method is significantly superior to DeepPPI. The Accuracy,

Precision, F_1 and AUC of MsaDtd are 10.43%, 8.38%, 4.35% and 7.48% higher than DeepPPI on CLNM dataset. They are increased by 5.38%, 6.53%, 3.8% and 2.06% comparing to DeepPPI on LLNM dataset.

5 Conclusion

In this paper, we propose an algorithm MsaDtd which improved decision tree with MS-Apriori and applied to the prognosis of thyroid disease through establishing a predictor to predict LNM in patients with PTMC. Fuzzy logic is introduced to handles continuous attributes, preventing to generate too many frequent items. Sorting and filtering rules mined by MS-Apriori used to avoid generate distractions, aiming to improve the prediction accuracy. Through the application of rules, the algorithm obtains new features to transform feature space, making full use of composite features. This improves the robustness and generalization capabilities of our algorithm. Building a decision tree and predicting thyroid disease by analyzing the affiliation between instances and rules to make the effective prediction. Clinicians can use the information given by predictor to adopt specific protocols throughout treatment. For the patients prone to LNM, clinicians should take customized interventions to reduce the risk of cancer recurrence.

Acknowledgement. Project supported by the Nature Science Foundation of Jilin Province (No. 20180101330JC), the National Nature Science Foundation of China (No. 60973040), the Fundamental Research Funds for the Central Universities (No. 2412017QD028), China Post-doctoral Science Foundation (No. 2017M621192), the Scientific and Technological Development Program of Jilin Province (No. 20180520022JH).

References

1. Fan, M., Hu, J., Cao, R., et al.: A review on experimental design for pollutants removal in water treatment with the aid of artificial intelligence. *Chemosphere* **200**, 330–343 (2018)
2. Jiang, F., Jiang, Y., Zhi, H., et al.: Artificial intelligence in healthcare: past, present and future. *Stroke Vasc. Neurol.* **2**(4), 230–243 (2017)
3. Jiang, H., Zhang, Z., Tao, L.: A semantic-based EMRs integration framework for diagnosis decision-making. In: Buchmann, R., Kifor, C.V., Yu, J. (eds.) *KSEM 2014. LNCS (LNAI)*, vol. 8793, pp. 380–387. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12096-6_34
4. Fang, R., Pouyanfar, S., Yang, Y., et al.: Computational health informatics in the big data age: a survey. *ACM Comput. Surv.* **49**(1), 12 (2016)
5. Vemulapalli, V., Qu, J., Garren, J.M., et al.: Non-obvious correlations to disease management unraveled by Bayesian artificial intelligence analyses of CMS data. *Artif. Intell. Med.* **74**, 1–8 (2016)
6. Tomaszewski, J.J., Uzzo, R.G., Egleston, B., et al.: Coupling of prostate and thyroid cancer diagnoses in the United States. *Ann. Surg. Oncol.* **22**(3), 1043–1049 (2015)
7. Akin, Ş., Yazgan, A.D., Akin, S., et al.: Prediction of central lymph node metastasis in patients with thyroid papillary microcarcinoma. *Turk. J. Med. Sci.* **47**(6), 1723 (2017)
8. Chen, H.L., Yang, B., Wang, G., et al.: A three-stage expert system based on support vector machines for thyroid disease diagnosis. *J. Med. Syst.* **36**(3), 1953–1963 (2012)

9. Makas, H., Yumusak, N.: A comprehensive study on thyroid diagnosis by neural networks and swarm intelligence. In: International Conference on Electronics, Computer and Computation, pp. 180–183. IEEE, Ankara (2014)
10. Pourahmad, S., Azad, M., Paydar, S.: Diagnosis of malignancy in thyroid tumors by multi-layer perceptron neural networks with different batch learning algorithms. *Glob. J. Health Sci.* **7**(6), 46–54 (2015)
11. Kaya, Y.A.: Fast intelligent diagnosis system for thyroid diseases based on extreme learning machine. *Arch. Otolaryngol. Head Neck Surg.* **15**(1), 41–49 (2014)
12. Maysanjaya, I.M.D., Nugroho, H.A., Setiawan, N.A.: A comparison of classification methods on diagnosis of thyroid diseases. In: International Seminar on Intelligent Technology and ITS Applications, pp. 89–92. IEEE, Surabaya (2015)
13. Chaudhary, R., Sharma, S., Sharma, V.K.: Improving the performance of MS-Apriori algorithm using dynamic matrix technique and map-reduce framework. *Int. J. Innov. Res. Sci. Technol.* **2**(5), 2349–6010 (2015)
14. Du, X., Sun, S., Hu, C., et al.: DeepPPI: boosting prediction of protein-protein interactions with deep neural networks. *J. Chem. Inf. Model.* **57**(6), 1499–1510 (2017)