Flu Diagnosis System Using Jaccard Index and Rough Set Approaches

Riswan Efendi\textsuperscript{1,\,*}, Noor Azah Samsudin\textsuperscript{1}, Mustafa Mat Deris\textsuperscript{1} and Yip Guan Ting\textsuperscript{1}

\textsuperscript{1}Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Malaysia
\textsuperscript{2}Mathematics Department, UIN Sultan Syarif Kasim, 28293 Pekanbaru, Indonesia

\*E-mail: riswan@uthm.edu.my

Abstract. Jaccard index and rough set approaches have been frequently implemented in decision support systems with various domain applications. Both approaches are appropriate to be considered for categorical data analysis. This paper presents the applications of sets operations for flu diagnosis systems based on two different approaches, such as, Jaccard index and rough set. These two different approaches are established using set operations concept, namely intersection and subset. The step-by-step procedure is demonstrated from each approach in diagnosing flu system. The similarity and dissimilarity indexes between conditional symptoms and decision are measured using Jaccard approach. Additionally, the rough set is used to build decision support rules. Moreover, the decision support rules are established using redundant data analysis and elimination of unclassified elements. A number data sets is considered to attempt the step-by-step procedure from each approach. The result has shown that rough set can be used to support Jaccard approaches in establishing decision support rules. Additionally, Jaccard index is better approach for investigating the worst condition of patients. While, the definitely and possibly patients with or without flu can be determined using rough set approach. The rules may improve the performance of medical diagnosis systems. Therefore, inexperienced doctors and patients are easier in preliminary flu diagnosis.

1. Introduction
In this paper, we are using two different approaches in diagnosing flu system, namely Jaccard index and rough set. All approaches is formed using set operations application, such as, intersection and subsets. In medical diagnostic domain, Venn diagram has been used to demonstrate the intersections between true positive, true negative, false positive and false negative results in a test for micro albuminuria [1]. While, Jaccard index has been also implemented to quantify similarity between hereditary diseases at molecular level [2]. However, the data reduction between conditional attributes and decision attribute cannot be solved using Venn and Jaccard approaches for medical diagnostic applications. Therefore, rough set theory was introduced and has been widely applied to solve complex problems by researchers in emergency room diagnostic medical. The rough set approaches can be used to assist such inexperienced doctors in diagnosing based on clinical decision support model of disease symptoms [3, 4].

There are existing rough set applications in medical diagnostic procedure to detect diseases, such as, dengue [3], diabetes mellitus [3, 4], chikungunya [4], and other. However, the step-by-step procedure...
in determining suitable rules for the medical diagnostic applications remains an interesting issue since the ultimate goal is to achieve accurate prediction results. Motivated by application of set theory and its operations in various medical diagnostic applications [1-4], we are interested to investigate the dependency of conditional attributes (flu’s symptoms) and decision attribute (flu) based on set operations from two different approaches. Furthermore, the symptom dependency values can be used for preliminary prediction purpose of the flu patients.

2. Fundamental Theories of Jaccard Index and Rough Set

In this paper, we explain two different approaches and its theories, namely, Jaccard index and rough set in determining the flu diagnosis system. While, both approaches and its theories will be explained in Sections 2.1 and 2.2.

2.1. Jaccard Index Theory

Motivated by Venn diagram and set operations, The Jaccard index, also known as intersection over union and the Jaccard similarity coefficient (originally coined coefficient de communauté by Paul Jaccard), is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets [5]:

\[ J(A, B) = \frac{|A \cap B|}{|A \cup B|} \]  

(1)

The Jaccard distance, which measures dissimilarity between sample sets, is complementary to the Jaccard coefficient and is obtained by subtracting the Jaccard coefficient from 1, or, equivalently, by dividing the difference of the sizes of the union and the intersection of two sets by the size of the union:

\[ dJ(A, B) = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \]  

(2)

2.2. Rough Set Theory

The rough set theory has been introduced by Pawlak (1982) and well divided by researchers into information systems, indiscernibility relation, set approximations, rough clustering, and others. An information system \( S = (U, \Omega, V_q, f_q) \) consists of \([6, 7, 8]\):

- \( U \) : a nonempty, finite set called the universe;
- \( \Omega \) : a nonempty, finite set of attributes;
- \( \Omega : C \cup D \), in which \( C \) is a finite set of condition attributes and \( D \) is a finite set of decision attributes; for each \( q \in \Omega, V_q \) is called the domain of \( q \);
- \( f_q \) : an information function \( f_q: U \to V_q \).

The cases, states, processes, patients, companies, and observations can be interpreted as objects or elements of rough sets. The attributes of each element can be assumed as symptoms, factors, and characteristic information. A relationship between conditional attributes and decision attribute can be explained using information table. In this table, the row and column correspond to objects and attributes, respectively. The starting point of rough set theory is the indiscernibility relation, generated by information about objects of interest.

Let \( S = (U, \Omega, V_q, f_q) \) be an information system, then any subset \( B \) of \( A \) determines a binary (equivalence) relation \( \text{IND}(B) \) on \( U \), which will be called \( B \)-indiscernibility relation, and is defined as follows:

\[ \text{IND}(B) = \{(x, y) \in U^2: \forall a \in B, a(x) = a(y)\} \]  

(3)

Where \( a(x) \) denotes the value of attribute \( a \) for element \( x \) in \( U \). The collection of all equivalence
classes determined by $\text{IND}(B)$, denoted by $U/B$. An equivalence class of $U/B$, containing $x$, is denoted by $[x]_B$.

Let $S = (U, \Omega, V_q, f_q)$ be an information system and let $B \subseteq A$ and $X \subseteq U$. We can approximate $X$ using only the information contained in $B$ by constructing the $B$-lower and $B$-upper approximations of $X$. Both approximations are denoted as:

$$B(X) = \{x \in U | [x]_B \subseteq X\},$$

(4)

And

$$\bar{B}(X) = \{x \in U | [x]_B \cap X \neq \emptyset\},$$

(5)

Where $[x]_B$ is an equivalence class containing $x$. While, the difference between both approximations and its accuracy can be written:

$$\text{BND}(X) = B(X) - \bar{B}(X),$$

(6)

3. Implementation

In this section, we discuss how to build the flu diagnosis system using Jaccard index and rough set approaches in sections 3.1 and 3.2.

3.1. Flu Diagnosis Using Jaccard Index Approach

This sub-section presents the implementation of set operations using Jaccard index in diagnosing flu system. A number data sets of patient flu [9] is presented in Table 1.

**Table 1. Information of patient’s flu and its symptoms**

| Patient’s code | Conditional attributes | Decision attribute |
|----------------|------------------------|--------------------|
|                | Headache | Muscle Pain | Temperature | Flu |
| $p_1$          | No       | Yes         | High        | Yes |
| $p_2$          | Yes      | No          | High        | Yes |
| $p_3$          | Yes      | Yes         | Very High   | Yes |
| $p_4$          | No       | Yes         | Normal      | No  |
| $p_5$          | Yes      | No          | High        | No  |
| $p_6$          | No       | Yes         | Very High   | Yes |

Step 1: Based on table 1, define sets of conditional attributes and decision attribute using patient’s code as presented in table 2.

**Table 2. Sets attributes and its elements**

| Set Attributes | Set elements |
|----------------|--------------|
| Headache $= H$ | $\{\{\text{Yes}\}, \{\text{No}\}\}$, |
|                | $\{\{p_2, p_3, p_5\}, \{p_1, p_4, p_6\}\}$. |
| Muscle Pain $= MP$ | $\{\{\text{Yes}\}, \{\text{No}\}\}$, |
|                | $\{\{p_2, p_3, p_5\}, \{p_1, p_4, p_6\}\}$. |
| Temperature $= T$ | $\{\{\text{Normal}\}, \{\text{High}\}, \{\text{Very High}\}$, |
|                | $\{\{p_1\}, \{p_2, p_5\}, \{p_3, p_6\}\}$. |
| Decision (Flu) $= F$ | $\{\{\text{Yes}, \{\text{No}\}\}$, |
|                | $\{\{p_1, p_2, p_3, p_4, p_5\}\}$. |

Step 2: Find set intersections between conditional attributes and decision attribute.

Intersection between sets $H$ (Yes) and $MP$ (Yes), $H = \{p_2, p_3, p_5\}$, and $MP = \{p_1, p_3, p_4, p_6\}$. Then, $H \cap MP = \{p_3\}$.

Intersection between sets $H$ (Yes), $MP$ (Yes) and $T$ (very high), respectively. $H = \{p_2, p_3, p_5\}$, $MP = \{p_1, p_3, p_4, p_6\}$, $T = \{p_3, p_6\}$. Then, $H \cap MP \cap T = \{p_3\}$. 

3
Intersection sets $H$ (Yes), $MP$ (Yes), $T$ (Very high) and $F$ (Yes), respectively. $H = \{p_2, p_3, p_4\}$, $MP = \{p_1, p_5, p_6\}$, $T = \{p_3, p_6\}$, $F = \{p_1, p_2, p_3, p_6\}$. Then, $H \cap MP \cap T \cap F = \{p_3\}$. It is presented in figure 1.

![Figure 1. Intersections of sets H, MP, T and F](image)

Step 3: Based on Step 2, calculate Jaccard similarity index ($J$) for each figure as presented in table 3.

| Intersection of sets | Jaccard similarity and dissimilarity indexes ($J$) |
|----------------------|--------------------------------------------------|
| $H \cap MP = \{p_3\}$ | $= \frac{1}{6}$, $= 0.1667 = 16.67\%$, Jaccard similarity index is: $J_s(H, MP) = 16.67\%$ similar. Jaccard dissimilarity index is: $= 1 - J(H, MP)$, $= 1 - 0.1667$, $= 0.8333$. $J_d(H, MP) = 83.33\%$ dissimilar. |
| $H \cap MP \cap T = \{p_3\}$ | $= \frac{1}{7}$, $= 0.1429 = 14.29\%$, Jaccard similarity index is: $J_s(H, MP, T) = 14.29\%$ similar. Jaccard dissimilarity index is: $= 1 - J(H, MP, T)$, $= 1 - 0.1429$, $= 0.8571$. $J_d(H, MP, T) = 85.71\%$ dissimilar |
| $H \cap MP \cap T \cap F = \{p_3\}$ | $= \frac{1}{9}$, $= 0.1111 = 11.11\%$, Jaccard similarity index is: $J_s(H, MP, T, F) = 14.29\%$ similar. Jaccard dissimilarity index is: $= 1 - J(H, MP, T, F)$, $= 1 - 0.1111$, $= 0.8889$. $J_d(H, MP, T, F) = 88.89\%$ dissimilar |

Step 4: Based on step 3 and table 3, determine the definitely patient have flu. In this case, we obtain that the patient $p_3$ is the worst condition based on similarity and distance of $J_s(H, MP)$, $J_s(H, MP, T)$, and $J_s(H, MP, T, F)$. Therefore, the patient $p_3$ should be treated immediately.

3.2. Flu Diagnosis Using Rough Set Approach
Based on tables 1, 2 and [9], the rough set can be implemented to determine the dependency between conditional symptoms and decision attribute by following

Step 1: Determine lower, upper approximations and boundary regions as shown in tables 4 and 5.

**Table 4. Lower and upper approximations**

| Lower approximation (LA) | Upper approximation (UA) |
|--------------------------|--------------------------|
| The patients that are definitely have flu = \{p1, p3, p6\} | The patients that possibly have flu = \{p1, p2, p3, p6\}. |
| The patients does not have flu = \{p4\}. | The patient that possible does not have flu = \{p4, p5\}. |

**Table 5. Boundary regions (BR)**

| BR for definitely have flu | BR for possibly have flu |
|---------------------------|-------------------------|
| BR = \{ p1, p3, p6\} - \{ p1, p2, p3, p6\} | BR = \{ p4\} - \{ p4, p5\} |
| = \{p2\}. | = \{p5\}. |

By following all steps given in data reduction and extraction [9], we only present the final result of intersection data with symptoms and decision attributes on table 6.

**Table 6. Final intersection data and information**

| Patient code | Conditional attribute | Decision attribute |
|--------------|-----------------------|-------------------|
|              | Headache | Muscle Pain | Temperature | Flu |
| p3           | Yes      | Yes         | Very High   | Yes |
| p6           | No       | Yes         | Very High   | Yes |
| p1           | No       | Yes         | High        | Yes |
| p4           | No       | Yes         | Normal      | No  |

Based on table 6, we generate the decision support rules for flu diagnosis (prediction) as presented in table 7, and we can also find the final intersection and information in table 8:

**Table 7. Proposed decision support rules**

| Rule | Condition |
|------|-----------|
| Rule 1 | If a patient with Symptom Headache: “Yes”, and Symptom Muscle Pain: “Yes”, and Symptom Temperature: “Very High”. Then decision of flu: “Yes”. |
| Rule 2 | If a patient with Symptom Muscle Pain: “Yes”, and Symptom Temperature: “High or Very High”. Then decision of Flu: “Yes”. |
| Rule 3 | Otherwise, “No Flu”. |

**Table 8. Final intersection data and information**

| Patient code | Conditional attribute | Decision attribute |
|--------------|-----------------------|-------------------|
|              | Headache | Muscle Pain | Temperature | Prediction attribute |
| p1           | No       | Yes         | High        | Headache | Muscle Pain |
| p2           | Yes      | No          | High        | p1       | No         | Yes         |
| p3           | Yes      | Yes         | Very High   | p2       | Yes        | No          |
| p4           | No       | Yes         | Normal      | p3       | Yes        | Yes         |
| p5           | Yes      | No          | High        | p4       | No         | Yes         |
| p6           | No       | Yes         | Very High   | p5       | Yes        | No          |
Table 7 shows the implementation of proposed rules in flu predicting and comparison result with rules proposed in [9]. Our proposed rules are able to predict almost all patients correctly, except the patient p2. While, the previous rule [9] is only able to predict the patient p1. In this case, the refining of rules are very important to be considered in order to improve the prediction accuracy.

4. Conclusion
In this paper, we implemented Jaccard index and rough set approximation for medical diagnosis systems, namely, flu diagnosis system. In the application, Jaccard index approach can be used to determine the worst condition based on the similarity and distance between conditional attributes and decision attribute. While, the rough set approach can be applied to determine the data reduction and decision support rules. Both approaches can be implemented to build the decision support systems for medical diagnosis. Thus, the inexperience doctors may use the systems for preliminary diagnosis of the patients.

Acknowledgement
This paper work is financially supported by ORICC Vot number D003 under Universiti Tun Hussein Onn Malaysia (UTHM).

References
[1] Stralen K J V, Stel V S, Reitsma J B, Dekker F W, Zoccali C, Jager K J 2009 Kidney Inter. 75 1257
[2] Caniza H, Romero A E, Paccararo A 2015 Sci. Report 5 17658
[3] Ali R, Hussain J, Siddiqi M H, Hussain M, Lee S 2015 Sensors 15 15921
[4] Thivigar M L, Richard C, Paul N R 2012 Inter. J. Inform. Sci. 2 33
[5] Jaccard P 1912 New Phytologist 11 37
[6] Pawlak Z 1982 Inter. J. Inform. Comp. Sci. 11 341
[7] Hampton J. 1997 J. Comp. Intell. Fin. 5 25
[8] Tay F E H, Shen L 2002 Euro. J. Operation. Res. 141 641
[9] Rissino S, Torres G L. 2009 Data Mining Knowledge Dis. in Real Life App. Inform. 35