Application of C4.5 Algorithm for Cattle Disease Classification

Evta Indra*, Kevin Ho, Arlinanda, Ridho Hakim, Delima Sitanggang, Oloan Sihombing
Faculty of Technology and Computer Science, Universitas Prima Indonesia, Indonesia
E-mail: evtaindra@unprimdn.ac.id*

Abstract. Data mining is a series of process to gain values such as informations that is manually unknown from a database by mining patterns from the data (with the intention to manipulate data into more useful information that is obtained by extracting data and recognizing important pattern or pulling data from the database). The intend of this research is to process cattle disease cases, which so far no research has been done to produce useful knowledge/science for institutions (Food and Livestock Security Agency) using data mining technique (C4.5 Algorithm). In the future, this research will be used by Food and Livestock Security Agency to counsel rancher in some area. This research will result a decision tree which will also classify rate of cattle disease. Proven by using the C4.5 algorithm, it can speed up the process of classification disease which initially takes about 1-2 months, and after using this application the results can be obtained right away.

1. Introduction

The increasing demand to resolve the limitation in processing data that many institutions can’t conduct a research to produce a useful knowledge for the institution or company. One of which is the Food and Livestock Security Agency (North Sumatera), which is in charge of animal health. So far, the activity only recorded and saved by the institution which is caused by the inability to process the data. But with the growth in Information Technology (IT), many solutions is available to resolve the situation. One of which is using the data mining technique.

Data mining is a pattern searching process and hidden relations in a big number of data with the intention to classify, estimate, predict, rule association, clustering, description and visualization [1]. One of the techniques in data mining is classification technique. Classification is the process of finding a frame or functions that show and distinguish concepts and data class. The purpose of classification is to predict a class objects that has a label which is not yet detected in data into the detected label, and one of the classification techniques that can be used is the C4.5 Algorithm.

The utilization of C4.5 algorithm to produce the level of data accuracy has been absolutely proved in previous research. Those researches are C4.5 algorithm used to predict the student achievement. The result shows that the accuracy prediction get up to 80.84% so that effective model prediction achievement can help education management department in finding the student bad-behavior and give the student guidance [2]. Beside, there is another researches of C4.5 algorithm is used to classify the fish disease. And as a result, the assess of fish disease performance obtained where both the lowest class and the highest class accuracy are founded which is respectively at 55.3% and 88.4% [3]. The other researches is the decision tree of C4.5 algorithm use in criminal case is to classify the data of...
prisoners to know the factors of detainees committing criminal acts of drugs. The result show that the most influential factor of the detainee committed the criminal act of drugs was from the address variable [4].

Based on the issue above, the writer is interested to conduct a research titled “Application of C4.5 Algorithm for Cattle Disease Classification”. The results of this research are hopefully to find an accurate performance analysis output in researched the cattle disease. So the results can be used to predict upcoming cattle disease and also can be used as decision making material in medicine stocks.

2. Methods

2.1. Data Mining
Data Mining is one of the fastest growing fields due to the huge demands for added value of large-scale databases that are in line with the growth of information technology and can extract large data sets into new knowledge. And also processing one or more machine learning technique to analyze and extract automated work [5]. Another definition is a induction-based learning, a process of forming defining general concept which is done with observing specific example from the concepts to be learned. In this context, data mining is a step of the KDD process and currently receives great attention and is recognized as a newly emerging analysis tool [6][7]. Knowledge Discovery in Databases (KDD) is a scientific application method in data mining. These applications are found in fields such as artificial intelligence, machine learning, market analysis, statistics and database systems, business management and decision support [8].

2.2. Data Mining Characteristics
In the processing technique, data mining has some characteristics, as follows:
- Data mining is associated with finding hidden things and data patterns that is unknown before.
- Data mining usually uses a very big data. Usually, big data is used to make results more reliable.
- Data mining is useful to make critical decision.

Based on descriptions above, it can be summed up that data mining is processing bunch of data or big information which is stored in data warehouse (database) to produce new data pattern from unknown data before [9].

2.3. C4.5 Algorithm
C4.5 Algorithm was introduced by Quinlan for inducing classification models and also referred as decision tree which the data based on the provided training data by way gain ration [10]. Classification is a data mining techniques that can be used to predict group membership to a data instance [11]. C4.5 Decision Tree is the very first fundamental supervised machine learning classification algorithm which is extensively implemented and typically achieves very good performance in prediction [12]. A decision tree is a flow-chart-like tree structure that can be constructed from the given set of attributes, where each branches represents an out-come of the test, and each leaf node represent classes [13][14]. C4.5 Algorithm is an expansion from ID3 [15]. Speed of C4.5 is significantly faster than ID3 (it is faster in several orders of magnitude) and C4.5 is more memory efficient than ID3 [16]. The resulting decision tree is the result of the algorithm C4.5 and can represent and model the results of the exploration of significant data, so the knowledge or information from these data more easily identified [17]. Some development which has been done to C4.5 is can overcome missing value, overcome continue data and pruning.

Generally, using C4.5 algorithm to build a decision tree is as follows:
- Select attribute as root.
- Make branches for each root.
- Divide cases into the branch.
- Repeat the process to every branch until all cases in the branches have the same class.
To choose attribute as root, based on the highest gain value from the existing attributes.

To calculate the gain, use the formula in the equation below:

\[
\text{Gain}(S,A) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} \text{Entropy}(S_i)
\]

With: S as set of cases; A as attribute; n as number of partitions in A attribute; |S_i| as number of i partitions; |S| as set of cases in S

Meanwhile, entropy value can be seen in the equation 2 below:

\[
\text{Entropy}(S) = \sum_{i=1}^{n} -p_i \log_2 p_i
\]

With: S as set of cases; A as feature; n as number of partitions S; p_i as proportion of S_i against S

There are 3 kinds of nodes in decision tree, as follows: Root node, internal node, and leaf node. The first is root node, is the topmost node, this node has no input and can’t have any output or having more than one output. The second is internal node, is the branching node, this node can only have one input and two minimal outputs. And the third is Leaf node or terminal node, is the last node, this node can only have one input and have no output.

In the Figure 1, for the class diagram of cattle disease diagnosis, there are 5 class as follows, Disease Table, Cases Table, Cases Data Relation Table, Reports Table, and Excel Table. This class diagram illustrates the system prototype / blueprint of how the system will be created.

| Diseases Table | Cases Table |
|----------------|-------------|
| DiseaseCode : String | DiseaseCode : String |
| Diagnosis : String | Diagnosis : String |
| Sign : String | Sign : String |
| Description : String | Description : String |
| +Add() | +Add() |
| +Edit() | +Edit() |
| +Cancel() | +Cancel() |
| +Delete() | +Delete() |
| +Exit() | +Exit() |

| Cases Data Relation Table |
|---------------------------|
| IdCase : String |
| DiseaseCode : String |
| Total : Int |
| +Add() |
| +Edit() |
| +Cancel() |
| +Delete() |
| +Exit() |

| Reports Table | Excel Table |
|---------------|-------------|
| DiseaseCode : String | DiseaseCode : String |
| Diagnosis : String | Diagnosis : String |
| Sign : String | Sign : String |
| Description : String | Description : String |
| +Add() | +Add() |
| +Delete() | +Delete() |

**Figure 1.** Class Diagram

Processing only use three attributes: diagnosis, sign and number of cases, so data transformation to the will-be-processed data is necessary. After data transformation has been done, then later, process will be done with C4.5 Algorithm. Sample of the data transformed can be seen in the Table 1:

**Table 1.** Disease Data Transformation (this data is taken from Food and Livestock Security Agency)

| No | Signs             | Temporary Diagnosis         | Infected Cattle |
|----|-------------------|------------------------------|-----------------|
| 1  | Placental Retention | Retensio Secundinarum        | 23              |
| 2  | Running Nose, Limping | Bovine Ephemeral Fever      | 12              |
| 3  | Fever              | Bovine Ephemeral Fever       | 98              |
|   | Anorexia, Running Nose | Bovine Ephemeral Fever |   |
|---|------------------------|------------------------|---|
| 4 | Limping, Bleeding Wounds, Ulcers | Accident | 71 |
| 5 | Diarrhea, Thinnes | Worms | 135 |
| 6 | Dull Fur | Worms | 39 |
| 7 | Thinnes, Anorexia, Diarrhea | Worms | 11 |
| 8 | Itchy, Ulcers | Ring Worm | 1 |
| 9 | Uterine Prolapse | Uterine Prolapse | 2 |
| 10 | Diarrhea | Pesticide Poisoning | 7 |
| 11 | Birth Difficulties | Distoxia | 14 |
| 12 | Flesh grow | Papilloma | 7 |
| 13 | Ulcers | Miasis | 5 |
| 14 | Thinness, Anorexia, Hair fall | Fasciolosis | 13 |
| 15 | Thinness, Itchy, Hair fall | Scabies | 15 |
| 16 | Inflammation of the eye | Eye Inflammation | 7 |
| 17 | Diarrhea, Anorexia | Salmonellosis | 39 |
| 18 | Swollen Udder, Anorexia | Mastitis | 1 |
| 19 | Bleeding Wounds, Limping | Tetanus | 6 |
| 20 | Thinness | Coccidiosis | 5 |
| 21 | Disabled | Lack of Calcium | 2 |
| 22 | Leg Wounds | Abses | 1 |
| 23 | Vagina Prolapse | Vagina Prolapse | 1 |
| 24 | Bloated | Tympany | 10 |
| 25 | Diarrhea | Colibacillosis | 9 |
| 26 | Limping, Swollen Joints | Footroot | 2 |
| 27 | Dull fur, Thinness | Malnutrition | 1 |

3. Results & Discussions

3.1. Use Case Diagram
Identificaton information system needs at Food and Livestock Security Agency can be seen in the use case diagram in the Figure 2. In the system, there are user and two main use cases namely, cases and process C4.5. In cases main use case, there are two include, first is diseases data that extend save, edit, and delete disease data and the second include is cases data that extend save, edit, and delete case data. For C4.5 process main use case, the extend is search data, C4.5 data processing, and reports. User logs via login, then system validates the user, display diseases data, cases data, C4.5 data process and reports.
3.2. Activity Diagram
The activity process of C4.5 algorithm can be seen at Figure 3. In the main menu, user can input cases starting from save, edit, search and delete the will-be-processed data. Subsequently, system will display data that has been chosen by the use, then user will view the data displayed. After book data displayed, then user can determine the cluster and system will process the data randomly to be made as initial cluster value. After that, user choose C4.5 algorithm process button then system will do the C4.5 algorithm process. After the process is finished, system will display the result of the C4.5 algorithm calculation, and C4.5 algorithm process is finished.
Data processing starts with finding the total entropy from the whole attribute and then determine the highest gain value. To get the gain value in creating the decision tree, we first need to calculate the information value in bits from a collection of projects. The calculation form for the entropy is as follows:

**Table 2. Calculation Results for Entropy Value and Gain**

| ATTRIBUTE                            | CASES | HIGH | MIDDLE | LOW | ENTROPY | GAIN |
|--------------------------------------|-------|------|--------|-----|---------|------|
| TOTAL CASES                          | 28    | 5    | 1      | 22  | 0.888   | 0.196|
| DISEASE                              |       |      |        |     |         |      |
| RETENSIO SECUNDINARUM                | 1     | 0    | 1      | 0   | 0       | 0    |
| BOVINE EPHEMERAL FEVER               | 3     | 1    | 0      | 2   | 0.918   |      |
| ACCIDENT                             | 1     | 0    | 0      | 1   | 0       |      |
| WORMS                                | 3     | 2    | 0      | 1   | 0.918   |      |
| RING WORM                            | 1     | 0    | 0      | 1   | 0       |      |
| UTERINE PROLAPSE                     | 1     | 0    | 0      | 1   | 0       |      |
| PESTICIDE POISONING                  | 1     | 0    | 0      | 1   | 0       |      |
| DISTOXIA                             | 1     | 0    | 0      | 1   | 0       |      |
| PAPILLOMA                            | 1     | 0    | 0      | 1   | 0       |      |
| MIASIS                               | 1     | 0    | 0      | 1   | 0       |      |
From the result of the calculations above, it is clear that the highest gain value is the disease attribute by 0.196 so the disease attribute becomes the root node. The decision tree can be seen as at Figure 4.
In the following decision tree, it is described that there are several branches of the node and as a sample a part of the branch node is taken in worm disease. For details can be seen in Figure 5. The diagnosis is identified as the root node and has a branch node derivative which is worms disease. In the worms disease branch node, there are 3 parts of classification. For symptoms 1 (thinness, anorexia, diarrhea) is included in the low class, for symptoms 2 (dull fur) including in the high class and for symptoms 3 (diarrhea, thinness) including in the high class.

The following is a data processing form that can be seen at Figure 6. In C4.5 process form, there is a period (period) whose function is to determine the distance from the data to be predicted and after the period is determined, the data will appear in the training data Table. The process (proses) button
functions to calculate the C4.5 algorithm process from the database and to display the calculation of the C4.5 algorithm checklist the show calculation (tampilkan perhitungan) checkbox. Chart button is to display the graph of disease samples form. Print button to display the reports form. Prediction (prediksi) button to find out the level of cattle that are infected by the disease diagnosed.

![C4.5 Process Form](image)

**Figure 6. C4.5 Process Form**

In the Figure 7, there is a demonstration of the C4.5 algorithm process, for example the period is taken from may 2015. After that press the load button, then the data will be displayed in the training data list. Then, press the process button so that the program will do the C4.5 algorithm calculation process. And last, choose the type of disease that wants to be diagnosed, for example is worms (cacingan). Then, a notification will show the number of cattle that diagnose infected by worms disease is high.
In the Figure 8, there is the form that displays the calculation of C4.5 algorithm process.
In the Figure 9, there is the form that show the graph of disease samples based on period. For example, this is a sample graph from the may 2015 period and the diagnosis for the highest disease is worms.

![Figure 9. Graph of Disease Samples Form](image_url)

In the Table 3, there is the reports form. In this form show the results of the C4.5 algorithm calculation for the specific period. In this form there are 4 attributes, as follows: disease code (kode penyakit), diagnosis (diagnosa), sign (tanda), and description (keterangan).

| Disease Code | Diagnosis               | Sign             | Description |
|--------------|-------------------------|------------------|-------------|
| B-001        | Fasciolosis             | Kekurusan        | Rendah      |
| B-002        | Cacingan                | Kekurusan, Mencret | Tinggi   |
| B-003        | Prolap uteri            | Prolaps rahim    | Rendah      |
| B-004        | Prolap vagina           | Prolaps rahim    | Rendah      |
| B-007        | Bovine Ephemeral Fever  | Pincang, Demam  | Tinggi      |
| B-008        | Keceilakan              | Pincang          | Tinggi      |
| B-010        | Coccidiosis             | Mencret          | Rendah      |
| B-014        | Tympany                 | Kembung          | Rendah      |
| B-016        | Distoxia                | Kesulitan lahir  | Rendah      |
| B-017        | Retensio Secundinarum   | Retensi placenta | Sedang      |
| B-021        | Scabies                 | Gatal, Bulu rontok | Rendah  |
| B-022        | Miasis                  | Abses, Luka berdarah | Rendah  |
| B-023        | Keracunan Pestisida     | Mencret, Susah berdiri | Rendah  |
| B-025        | Radang mata             | Radang mata      | Rendah      |
| B-027        | Kekurangan Calsum       | Pincang          | Rendah      |
| B-028        | Tetanus                 | Luka berdarah    | Rendah      |
| B-029        | Mastitis                | Ambing bengkak, Kelainan ambing | Rendah  |
4. Conclusion
From the results of the data mining application to detect cattle disease, the conclusion is that the system can provide information needed for investigative management to classify cattle disease quickly, which initially takes about 1-2 months to classify cattle disease. After using this application, then the classification results can be obtained directly. Then, the supply of medicines will be more effective and will also reduce the risk of disease infection in cattle. Other than that, the results of this classification of cattle disease can also be used for counseling to farms in certain areas.

References
[1] Fiandra Y A, Defit S and Yuhandri 2017 Penerapan Algoritma C4.5 Untuk Klasifikasi Data Rekam Medis Berdasarkan International Classification Diseases (ICD-10). 1 82-89
[2] Li L, Yao S, Ou Z and Chen Q 2015 Forecast of Student Achievement Variation Trend Based on C4.5 Decision Tree. 383-86
[3] Sucipto, Kusrini and Taufiq E L 2016 Classification Method of Multi-class on C4.5 Algorithm for Fish Diseases. 5-9
[4] Wahyuni S 2017 Implementation of Data Mining to Analyze Drug Cases Using C4.5 Decision Tree. 970 1-6
[5] Arifin M F and Devi F 2017 Penerapan Algoritma Klasifikasi C4.5 dalam Rekomendasi Penerimaan Mitra Penjualan Studi Kasus: PT Atria Artha Persada 8 87-102
[6] Jantan H, Hamdan A R and Othman Z A 2010 Human Talent Prediction in HRM using C4.5 Classification Algorithm. 2 2526-34
[7] Widayu H, Nasution S D, Silalahi N and Mesran 2017 Data Mining Untuk Memprediksi Jenis Transaksi Nasabah Pada Koperasi Simpan Pinjam Dengan Algoritma C4.5. 1 32-37.
[8] Adhatura K, Gaykar A, Dhawan A, Jha R and Honrao V 2013. Predicting Students’ Performance Using ID3 And C4.5 Classification Algorithms. 3 39-52
[9] Yuliga M, Rusli M and Winarso E 2015 Prediksi Harga Emas Dunia Sebagai Pendukung Keputusan Investasi Saham Emas Menggunakan Teknik Data Mining. 36-51
[10] Meena P I M and Perumal V 2016 Performance of C4.5 and Naïve Bayes Algorithm to Predict Stomach Cancer - An analysis. 5 482-87
[11] Muslim M A, Herowati A J, Sugiharti E and Prasetyo B 2017 Application of the pessimistic pruning to increase the accuracy of C4.5 algorithm in diagnosing chronic kidney disease. 983 1-9
[12] Jothikumar R and Siva RVB 2016 C4.5 classification algorithm with back-track pruning for accurate prediction of heart disease. S107-11
[13] Agrawal G L and Gupta H 2013 Optimization of C4.5 Decision Tree Algorithm for Data Mining Application. 3 341-45

[14] Manikandan M and Mala R 2017 Optimal Prediction of Weather Condition Based on C4.5 Classification Technique. 7 265-72

[15] Kustiyaningsih Y and Rahmanita E 2016 Aplikasi Sistem Pendukung Keputusan Menggunakan Algoritma C4.5 Untuk Penjurusan SMA. 5 101-8

[16] Kathija A, Nisha S S and Sathik M M Classification Of Breast Cancer Data Using C4.5 Classifier Algorithm. 2 13-19

[17] Buulolo E Silalahi N, Fadlina and Rahim R 2017 C4.5 Algorithm To Predict the Impact of the Earthquake. 6 10-15

Acknowledgements
This research would not have been possible without the support of Food and Livestock Security Agency (North Sumatera) with the permission to use cattle disease sample data as training data needed in this study. We thank our colleagues from Universitas Prima Indonesia (UNPRI) who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper. Nobody has been more important to the writer in pursuit this project than the writer's family members. The writer would like to thank to the writer’s parents, whose love and guidance with the writer every time the writer pursue. They are the ultimate role models, and who provide unending inspiration.