International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 7 Issue 6 Ver I || June 2018 || PP 01-08

Predicting Top-k Keywords in Document Streams Using Machine Learning Techniques

[1]Dr.G.Anandharaj, [2]S.K.Thilagavathy
[1]Associate Professor & Head, [2]M.Phil (CS) Research Scholar
[1,2]Department of Computer Science and Applications,
[1,2]Adhiparasakthi College of Arts and Science (Autonomous),
G.B.Nagar, Kalavai -632506, Vellore (District)
Corresponding auther : Dr.G.Anandharaj

Abstract : *The large hierarchy of documents accessible on the online and increasing dramatically each day. This huge volume of largely for the most part unstructured text can't be simply handled and seen by servers. Therefore, practiced and viable procedures and algorithms are needed to get helpful patterns. Keyword mining is that the task of extracting significant info from Documents, that has gained important attentions in recent years. During this paper, we have a tendency to describe many of the foremost elementary techniques for Top-K Keyword for Document Streams. We have a tendency to utilize weka Tool 3.8 is a point of interest framework within the historical background of the data mining and machine learning analysis teams. In these we have a tendency to examines an algorithmic rule to exactly classify the whole stream in to a given variety of reciprocally exclusive together thorough streams are often run additional relevant results with a high potency. We've known an array of ways that may be applied like k-Nearest Neighbors (kNN), Support Vector Machine (SVM) algorithms, and two trees based mostly classification algorithms: Random Forest and J48. J48 is that the Java implementation of the algorithmic rule C4.5. Algorithmic rule within which every node represent one among the possible selections to be taken and every leave represent the expected category. This paper describes the usage of machine learning techniques to assign keywords to documents.*

Keywords: -*Machine Learning; Keyword search, Information extraction; Clustering and classification*

Date of Submission: 18-05-2018 Date of acceptance: 04-06-2018

I. INTRODUCTION

This paper describes the usage of machine learning techniques to assign keywords to documents. The large hierarchy of documents accessible on cyber web, the Yahoo hierarchy, is employed here as a real-world downside domain. Machine learning techniques developed for learning on text info areused here inside the stratified classification structure. The high variety of choices is reduced by taking into account the information structure and using a feature set alternative supported the plan of action used in data retrieval. Documents are pictured as word-vectors that embody word sequences (tag-words) instead of merely single words. The info structure of the examples and class values is taken into account once shaping the sub problems and forming coaching job examples for them. Additionally, a data structure of sophistication values is used in classification, where exclusively promising strategies within the hierarchy are thought of. Text classification is used in information extraction and retrieval from a given text, and text classification has been thought of as an important step to manage an enormous varies of records given in digital kind that is sweeping and increasing. This thesis addresses patent document classification downside into fifteen all completely different categories or classes, where some classes overlap with various classes for wise reasons. For the event of the classification model victimization machine learning techniques, useful choices ar extracted from the given documents. The choices ar used to classify patent document what is more on generate useful tag-words. This can be seen as further information concerning the documents or as a style of document abstraction. For example, many conferences would like that each paper submission is within the course of a page containing a bunch of keywords describing the planet to be mentioned. Typically the authors are asked to choose keywords from the planned set of keywords given inside the conferences need papers. Here we are going to describe the usage of machine learning techniques for the matter of mechanically assignment keywords to documents. Our set of keywords is made public by the domain that is throughout this case the Yahoo hierarchy.

II. Literature Review

2.1 Review of Multi-Class Document Classification

In a pioneering approach, Fabrizio Sebastiani mechanized machine-controlled classification (or classification) of texts into predefined categories, this approach had seen AN increasing enthusiasm in recent years, as a result of the swollen accessibility of documents in digital type and also the subsequent ought to classify them [8]. The predominant thanks to trot out these text-classification issues ar popularly and manageably done victimization machine learning techniques, a general inductive method that naturally builds a text classifier by learning from a collection of tagged text documents, the attributes of the classes. The advantages of this approach over the information engineering approach (manual classification and annotation by domain specialists) are an honest viability, and direct immovableness to numerous domains. The success story of machine-controlled text classification is increasing to neighboring fields of application. For instance, creaking text classification ensuing from the optimum character recognition and speech transcripts ar on the market and galvanizing [8]. Text classification has advanced from a minor analysis field in late '80s, into a totally bloomed investigation field, that has produces practiced, powerful, and usually helpful classifications with several application areas. This action has been supported 2 aspects of modernization: initial, the frequently increasing inclusion of the machine learning cluster in text order, that has recently led to the use of the most recent machine learning innovation in text classification applications, and second, the accessibility of ordinary benchmarks, [9], that has supported analysis by giving a setting during which distinct analysis endeavors may well be contrasted with one another, and during which the simplest techniques and calculations might emerge.

2.2 Review of approaches in document classification

Currently, text classification has become one among the key strategies to handle and organize text information [10]. Documents, that usually contain strings of characters, need to be reworked into aSuitable illustration that usable in learning algorithms and classification issues. Info retrieval analysis suggests that word stems work alright as illustration units resulting in attribute worth illustration of text that is employed in implementing this project. The word stem comes from the shape of a word by removing case and derived info [11]. as an example "machine", "machining" square measure all mapped to an equivalent stem "machine". These results in Associate in Nursing attribute worth illustration of the text information. for every distinctive word, its count ωi within the information corresponds to a options term frequency (ωi,d) in corresponding text document d. to beat the unnecessarily massive feature vectors, words square measure painted as options if they need occurred within the coaching information set sometimes a minimum of three times. Supported this illustration, scaling the scale of the feature vector with their various inverse document frequency (IDF, that is applied because the log inverse ofω_i) leads to an improved performance. IDF is calculated from the total number of training documents (n) and the document frequency of the particular word ω_i as show in eqn (1).

$$\mathrm{IDF}(\omega_i) = \log\left(\frac{n}{\mathrm{DF}(\omega_i)}\right) \qquad (1)$$

Here,$\mathrm{DF}(\omega_i) =$ is the number of those documents in the collection, which contain the term(ω_i). Based on the standard feature vector representation of the text data, it was argued in [3] that the support vector machines area unit a lot of acceptable for this kind of setting. completely different classification ways like Thomas Bayes, SVM, C4.5 and kNN were applied on the Reuters-21578 and Ohsumed corpus [2] among that, SVM was found to possess superior prediction with sizable performance gains.

2.3 Overcoming data imbalance issue

According to R. Longadge et al. [5], imbalance information set is that the greatest drawback in data processing [35]. The authors have planned three strategies for classification of imbalance information set that is split into three classes particularly recursive, information preprocessing and have choice approach. The paper demonstrates systematic study of those approaches and therefore the right direction for the category imbalance (class imbalance happens once one amongst the two categories having additional sample than alternative classes) drawback. Sampling strategies square measure employed in resolution the info imbalance issues that are thought as information preprocessing methodology. Under-sampling and oversampling square measure the two strategies represented during this paper. The authors provided the strategies to perform the feature choice and recursive approaches and complete that the informationpreprocessing methodology provides far better resolution than alternative strategies because it permits the addition of latest information into the prevailing information or the deletion of the redundant data that is of less importance. Under-sampling is associate economical strategy to traumatize category imbalance however the downside of underneath sampling is that it loses several potential information as a number of the info is deleted. So as to realize sensible prediction over minority category and avoid necessary info loss from the bulk category, each k-means algorithmic rule and sampling approaches were enforced to get balanced information sets.

III. Preliminary Work

3.1 Approaches to document classification using machine learning

In supervised approach, single-label documents are those which are categoryified into one class solely, multi-label documents are those that are categoryified into quite one class [2]. During this project we tend to perform multi category classification, within which a file is foreseen into one in every of the predefined classes. Supervised learning models are wide applicable and might provide the insight concerning however the instructive variables ar associated with the specific response variable. Document classification chiefly consists of document illustration (vector form), feature choice, feature extraction, application of machine learning formula on the info for analysis. There are differing kinds of supervised text classification techniques on the market in machine learning platform, that we tend to applied to perform our preliminary assessment mentioned below: kNN (k Nearest Neighbors): kNN formula relies on the belief that the classification of a document is analogous to the classification of the opposite documents that ar close within the vector area. "k" in kNN could be a user-defined constant that is employed to search out the foremost frequent label of k coaching samples (documents) nearest to the unlabeled vector or check document. Support Vector Machines: it's a supervised classification formula that is extensively enforced in text classification issues. SVM have the potential to handle massive feature areas as they use over fitting protection that doesn't depend upon the quantity of attributes [37]. The task of SVM is to search out the linearly separators for the on the market classes.

Random Forest: it's a strong and versatile formula that is employed for classification ofdocuments. Random forest consists of the many individual trees [38]. During this technique, every tree votes on Associate in Nursing overall classification for the given dataset and therefore the random forest formula choses the foremost voted individual classification.

J48 call Trees: a choice tree could be a prognosticative machine learning model that decides the target price of a brand new sample supported many attribute values of the on the market coaching dataset. A binary tree is formed supported the coaching dataset and it's applied to focus on instance within the dataset for the classification. we tend to applied 5 machine learning ways on the experimental knowledge [3] to hold out the initial classification. These 5 ways are k-Nearest Neighbors (kNN), 2 variations of Support Vector Machine (SVM), and 2 tree-primarily based classification algorithms Random Forest and J48. We've assessed the accuracies of the initial classification with five fold cross-validation.

3.2 A Dissection of Support Vector Machine in Classification

Support Vector Machines may be a classifier formally outlined by a separating hyperplane in different words, given labeled coaching information (supervised learning), the algorithmic rule outputs classes. SVM may be a comparatively new category of machine learning techniques initial introduced by Vapnik [37] and has been introduced in Text classification by Joachims. The Support vector machine classifier wants each positive and negative coaching set, that area unit uncommon for different classification issues. This positive and negative coaching set area unit required for the SVM to hunt for the choice boundary that separates the 2 categories within the n dimensional house, therefore referred to as the hyperplane. The documents, that area unit nearest to the choice surface, area unit referred to as the support vectors. The accuracy or potency of the Support vector machine classifier classification doesn't alter if documents that don't belong to the support vectors area unit off from the set of coaching information.

Usually the classification drawback is confined to thought of thought of the two-class (binary) drawback while not loss of generality. The most goal of the matter is to separate the 2 categories by a perform that is iatrogenic from out there examples. The goal is to supply a classifier which will work well on unseen examples. Think about the instance in Figure one. Within the figure there are a unit several potential linear classifiers which will separate the information points, however there's only 1 classifier that maximizes the margin (maximizes the gap between it and also the nearest information of the 2 classes). Such linear classifier is outlined because the optimum separating hyperplane. Intuitively, we'd expect this hyperplane to generalize well as hostile the opposite potential category boundaries.

3.3 Attribute Selection

Attribute choice plays a key role in knowledge preparation for coaching the classifier model. It's a method within which the most effective set of attributes is chosen from the dataset. It's wont to produce transforms of the dataset like rescaled attribute values into their constituent elements to create additional and helpful structure for the classifier models. Keeping unsuitable attributes within the coaching model may cause over fitting downside.

Tokenization: Text knowledge accessible consists of terms sorted into sentences and paragraphs. One technique to represent text knowledge mistreatment the vector house model breaks down the matter knowledge into a group of terms. Every of those terms corresponds to associate degree attribute of the input file and so becomes associate degree axis within the vector house. The info is then delineated as vectors, whose parts

correspond to the terms contained within the knowledge assortment and whose worth indicates either a binary present/absent worth or a weightage for the term for the info purpose. This illustration is thought because the Extracted Term' illustration and is that the most typically used illustration for text knowledge patterns. The method of breaking the accessible stream of text into words is termed tokenization.

Feature extraction mistreatment String to Word Vector: This filtering methodology transforms string attributes gift within the text document into word vectors i.e., creates one attribute for every term that's gift within the document. The filtering choices for applying on the dataset are mentioned in section three.

Information Gain Attribute authority (InfoGainAttributeEval): Once the feature extraction is completed, feature choice is enforced bymistreatment the Attribute choice filter. info Gain Attribute authority gift within the filter permits USA to line the brink of the weights of the options to be selected for classification. In our project we've got set the brink to zero.0 specified all the options with positive info gain ar taken into thought.

3.4 A Machine Learning Toolkit, WEKA

Weka tool could be a assortment of machine learning algorithms for the info mining/document classification tasks. The algorithms can be applied to the datasets directly or known as from user developed java code. This application contains knowledge pre-processing, classification, regression,clustering etc. and plenty of alternative tools. It's well-suited for developing machine learning schemes and applications. [39] The subsequent choices that are accessible in StringToWordVector filter are used for preprocessing the coaching knowledge set. The tactic executes wont to normalize the reiterations to urge the multi-terms extraction.

The following choices are applied on StringToWordVec filter:

wordsToKeep - the quantity of words (per category if there's a category attribute assigned) to try to stay.

outputWordCounts - Output word counts instead of Boolean zero or one (indicating presence or absence of a word).

lowerCaseTokens - Set to create all the word tokens to regenerate to character before being supplemental to the lexicon.

normalizeDocLength - Sets the word frequencies for a document (instance) to be normalized.

TFTransform – TF stands for Term Frequency.

Applying Snowball Stemmer: This formula reduces inflected/derived words to their word stem, by trimming the foundation of its inflectional affixes; sometimes, solely suffixes that are supplemental to the right-hand finish of the foundation word ar removed. It reduces the repetitions of words of a similar kind that step by step reduces the spatiality of the word vectors.

Removing Stop words mistreatment Rainbow list: Stop words, are the words that are filtered out before the process of tongue knowledge (text). The presence of those words as attributes doesn't have any positive impact on the classification. They're sometimes outlined as functional-words, that don't carry that means and influence the classification completely. In our setup, we have a tendency to use rainbow stop words list and haveappended them with further words to create the list effective. So by the removal of those stop words, solely terms with high info gain can be thought of for the coaching purpose.

IV. Methods and their Integrations

Our experimental setup consists of the following 4 stages and the information flow has been shown by Figure 1:

i) Experimental Data Collection
ii) Data Preprocessing
iii) Applying Machine Learning algorithms
iv) Synthetic data Injection

The Experimental information is collected and preprocessed mistreatment StringToWordVec andAttributeSelection filters. Once the info is preprocessed, the model is trained with the info.

Depending on the accuracy of the model obtained mistreatment cross fold validation, artificial information is additional to the specified categories and also the method is recurrent from stage ii.

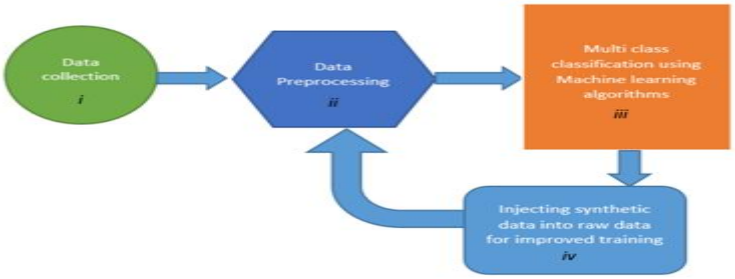


Fig 1: Text classification process

Stage 1: Experimental Data Collection
Stage 2: Data Preprocessing
Stage 3: Applying Machine Learning algorithms for classification
Stage 4: Addition of Synthetic in training data
The information gain of a feature F is the expected reduction in entropy resulting from splitting on this feature.

$$Gain(S,F) = Entropy(S) - \sum_{v \in values(F)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Where Sv is the subset of S having value v for feature F.
Entropy of each resulting subset weighted by its relative size.

V. Implementation

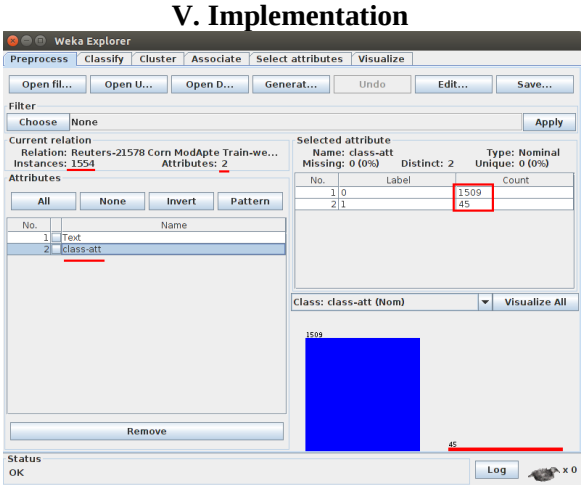


Fig 2. WEKA input files initial inspection

Shown in the illustration above is the number of instances in the file, the number of attributes for each instances, the distribution of corn and non-corn instances which is 45 (2.9%) corn and 1509 (97.1%) non-corn news.

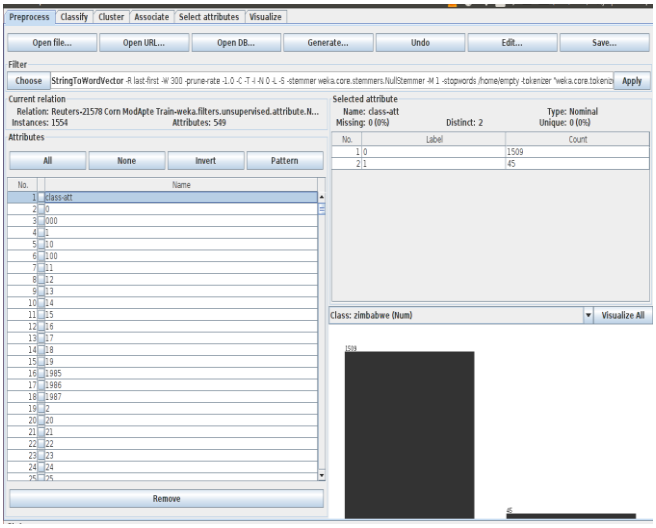


Fig. 3 Attributes list after feature extraction

Now you can see in the illustration above, the STWV produced more than 500 attributes (Features) where many of them are irrelevant.

You can use the Associator tab—the J48" algorithm- to discover relations between important attributes (you will need to use Filtered Associator → Discretize filter to be able to use it on this dataset)

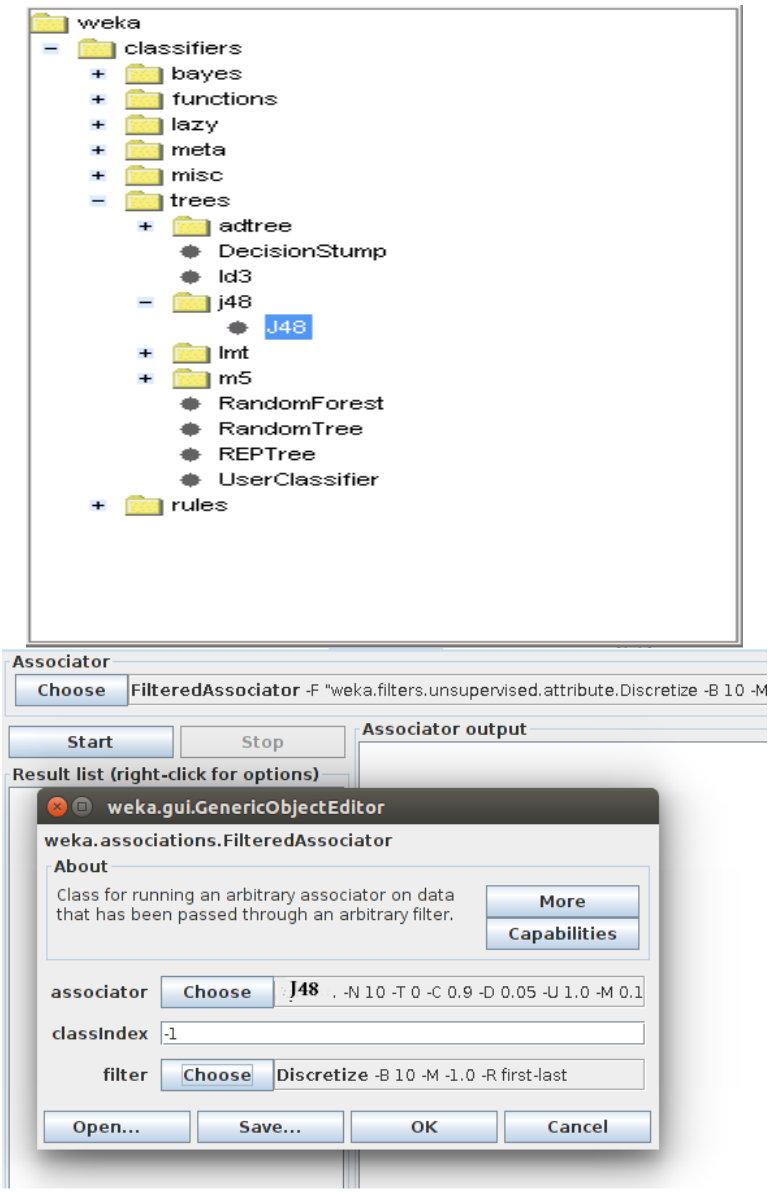


Fig. 4 A & 4 B Algorithm Selection

```
Best rules found:

 1. class-att=0 1509 ==> maize='(-inf-0.930853]' 1509    conf:(1)
 2. class-att=0 1509 ==> sorghum='(-inf-1.15775]' 1509    conf:(1)
 3. class-att=0 sorghum='(-inf-1.15775]' 1509 ==> maize='(-inf-0.930853]' 150
 4. class-att=0 maize='(-inf-0.930853]' 1509 ==> sorghum='(-inf-1.15775]' 150
 5. class-att=0 1509 ==> maize='(-inf-0.930853]' sorghum='(-inf-1.15775]' 150
 6. corn='(-inf-1.001058]' soybean='(-inf-1.053663]' 1508 ==> sorghum='(-inf-
 7. class-att=0 bushel='(-inf-1.119904]' 1506 ==> maize='(-inf-0.930853]' 150
 8. class-att=0 bushel='(-inf-1.119904]' 1506 ==> sorghum='(-inf-1.15775]' 15
 9. corn='(-inf-1.001058]' maize='(-inf-0.930853]' 1506 ==> sorghum='(-inf-1.
10. class-att=0 bushel='(-inf-1.119904]' sorghum='(-inf-1.15775]' 1506 ==> ma
```

Fig. 5 Best Rules

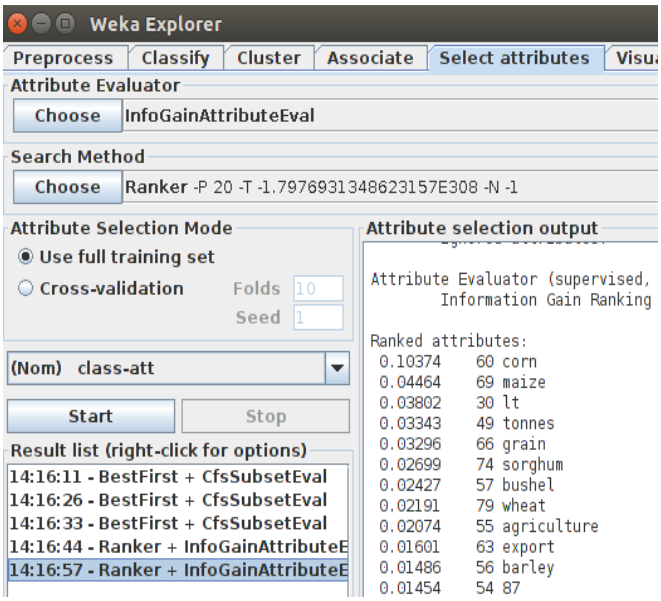


Fig. 6 Select Attributes

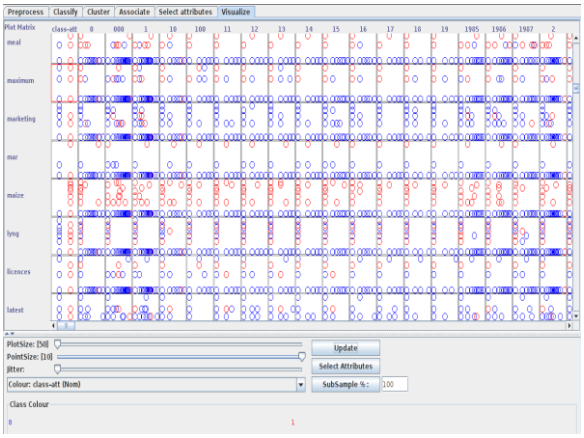


Fig.7 Visuvalization

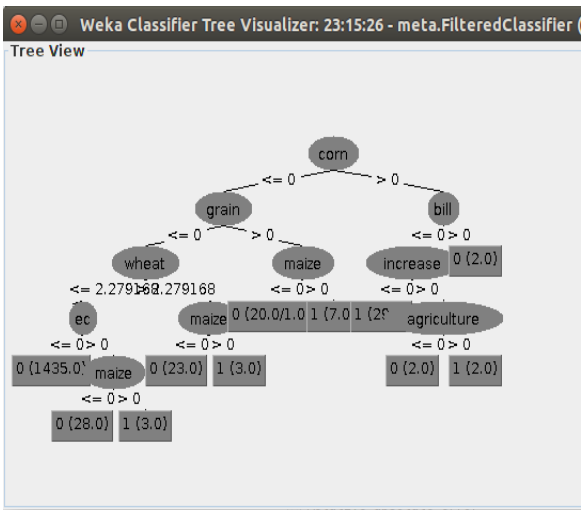
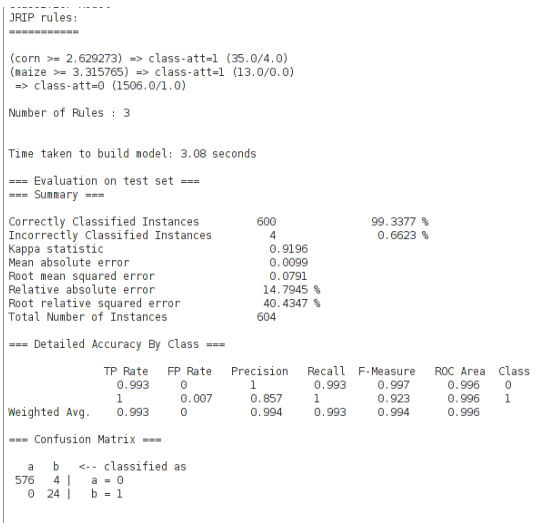


Fig. 8 Decision Tree



```
JRIP rules:
===========

(corn >= 2.629273) => class-att=1 (35.0/4.0)
(maize >= 3.315765) => class-att=1 (13.0/0.0)
 => class-att=0 (1506.0/1.0)

Number of Rules : 3


Time taken to build model: 3.08 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         600               99.3377 %
Incorrectly Classified Instances         4                0.6623 %
Kappa statistic                          0.9196
Mean absolute error                      0.0099
Root mean squared error                  0.0791
Relative absolute error                 14.7945 %
Root relative squared error             40.4347 %
Total Number of Instances              604

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.993     0          1         0.993     0.997      0.996    0
                 1         0.007      0.857     1         0.923      0.996    1
Weighted Avg.    0.993     0          0.994     0.993     0.994      0.996

=== Confusion Matrix ===

   a   b   <-- classified as
 576   4 |   a = 0
   0  24 |   b = 1
```

Fig. 9 Final Result

VI. Conclusion

For the event of the classification model victimization machine learning techniques, helpful options are extracted from the given documents. We've got known AN array of strategies that may be applied like k-Nearest Neighbors (kNN), 2 variations of the Support Vector Machine (SVM) algorithms, and 2 trees primarily based classification algorithms: Random Forest and J48. The foremost analysis steps during this work encompass filtering techniques for variable choice, data gain and have correlation analysis, and coaching and testing potential models victimization effective classifiers. Further, the obstacles related to the unbalanced knowledge were mitigated by adding artificial knowledge where acceptable, that resulted in a very superior SVM classifier primarily based model.

REFERENCES

[1]. [1] P. Haghani, S. Michel, and K. Aberer, "The gist of everything new: personalized top-k processing over web 2.0 streams." in CIKM, 2010, pp. 489–498.

[2]. K. Mouratidis and H. Pang, "Efficient evaluation of continuous text search queries," IEEE Trans. Knowl. Data Eng., vol. 23, no. 10, pp. 469–1482, 2011.

[3]. N. Vouzoukidou, B. Amann, and V. Christophides, "Processing continuous text queries featuring non-homogeneous scoring functions." in CIKM, 2012, pp. 1065–1074.

[4]. A. Hoppe, "Automatic ontology-based user profile learning from heterogeneous web resources in a big data context." PVLDB, pp. 1428–433, 2013.

[5]. A. Lacerda and N. Ziviani, "Building user profiles to improve user experience in recommender systems," in WSDM, 2013, pp. 759–764.

[6]. M. Busch, K. Gade, B. Larson, P. Lok, S. Luckenbill, and J. J. Lin,"Earlybird: Real-time search at twitter," in ICDE, 2012, pp. 1360– 1369.

[7]. L. Wu, W. Lin, X. Xiao, and Y. Xu, "LSII: an indexing structure for exact real-time search on microblogs," in ICDE, 2013, pp. 482–493.

[8]. J. Zobel and A. Moffat, "Inverted files for text search engines," ACM Comput. Surv., vol. 38, no. 2, 2006.

[9]. R. Fagin, A. Lotem, and M. Naor, "Optimal aggregation algorithms for middleware," J. Comput. Syst. Sci., vol. 66, no. 4, pp. 614–656, 2003.

[10]. A. Z. Broder, D. Carmel, M. Herscovici, A. Soffer, and J. Y. Zien, "Efficient query evaluation using a two level retrieval process." in CIKM, 2003, pp. 426–434.

Dr.G.Anandharaj "Predicting Top-k Keywords in Document Streams Using Machine Learning Techniques "International Journal of Engineering Science Invention (IJESI), vol. 07, no. 06, 2018, pp 01-08