Journal of Theoretical and Applied Information Technology
15th December 2016. Vol.94. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MODEL OF SECURITY LEVEL CLASSIFICATION FOR DATA IN HYBRID CLOUD COMPUTING

[1,2] MOHANAAD SHAKIR, [3] ASMIDAR ABUBAKAR, [3]YOUNUS YOUSOFF, [1]MOHAMMED WASEEM,[4] MOSTAFA AL-EMRAN

[1] Senior Lecturer, IT Department , Alburaimi University Collage(BUC), Oman
[2] Ph.D. Candidate in ICT, COGS, University Tenaga National(UNITEN), Malaysia
[3] Senior Lecturer, COIT, University Tenaga National(UNITEN), Malaysia
[4]Ph.D. Candidate in FSKKP, University Malaysia Pahang(UMP), Malaysia

E-mail: mohanaad@buc.edu.om , asmidar@uniten.edu.my, yunusy@uniten.edu.my, ashfaque@buc.edu.om, malemran@buc.edu.om

ABSTRACT

Organizations mainly rely on data and the mechanism of dealing with that data on cloud computing. Data in an organization has multi security levels, which is classified depending on nature of the data, and the impact of data on the organization. The security procedures which used for protecting data usually be complicated, and it had a direct and indirect influence on the usability level. This study aims to establish a model which has an ability to classify data dynamically according to the security form low till high levels. The security level classified it into five levels based on the policies and classification method. The purpose of classification is to apply a complex security procedure on data which has a high security level larger than data which has a low security level. It also has a potential to segregation an illegal data from the legal to support usability in system. Finally, several experiments have been conducted to evaluate the proposed approaches. Several experiments have been performed to empirically evaluate two feature selection methods (Chi-square (χ2), information gain (IG)) and five classification methods (decision tree classifier, Support Vector Machine (SVM), Naïve Bayes (NB), and K-Nearest Neighbor (KNN) and meta-classifier combination) for Legal Documents Filtering The results show that all classifiers perform better with the information gain feature selection methods than their results with Chi-Square feature selection method. Results also show that Support Vector Machine (SVM) outperforms achieve the best results among all individual classifiers. However, the proposed meta-classifiers method achieves the best results among all classification approaches.

Keywords: *Information System Security, Classification Of Data, Big Data, Neural Language Processing*

1. INTRODUCTION

Now a day's security issue became challenge for IT based infrastructure organizations, as a part of security well defined security level contents and rules are preferable than in the past. Each piece of information has its own security level; correct detection of this security level may lead to apply correct protection rules on information. Generally, organizations or institutions are very concerned in improving the security of cloud computing through the application of the authority model and dynamic classification of data model based on the multi-level security [1,2]. The aim of classification of confidential documents is to assign predefined class labels to a new document that is not classified [3]. An associated classification model provides training documents with existing class labels. Therefore, supervised, semi-supervised or unsupervised classification algorithms are fitting as a solution to the classification problem. The set of labeled and unlabeled documents for an organization may lead selecting supervised or unsupervised algorithms. For a document set which contains mostly unlabeled documents, choosing an appropriate unsupervised or semi-supervised methodology may present more accurate results. In the other hand if all the train and test documents are labeled by a subject matter expert, and then by using a supervised algorithm seems obviously more realistic procedure [4]. Classification accuracy of textual data is highly related to preprocessing tasks of training and test data [5]. These tasks become more difficult in processing unstructured textual data than in structured data. Unstructured nature of data needs to be formatted in a relational and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

analytical form. TF-IDF (term frequency-inverse document frequency) is preferred to represent text based contents of documents. This representation holds each word stem as an attribute for classification; and each document represents a separated classification event.

2. CLASSIFICATION COMPONENTS

These classifiers include decision tree classifier, Support Vector Machine (SVM), Naïve Bayes (NB), and K-Nearest Neighbor (KNN).

2.1 Decision Tree Learning (Dt) Classifier

Decision trees normally used as classifiers, i.e. they assign classes to objects which are represented as attribute vectors [6,7]. The non-terminal nodes are labeled with attribute tests, the edges with the possible outcomes of a test, and the terminal nodes are labeled with classes. An object is classified by evaluating the test of the top node on the object, following the respective edge to a daughter node, evaluating the test of the daughter node, and so on until a terminal node is reached whose class is assigned to the object. Decision Trees are turned into probability estimation trees by storing a probability for each possible class at the terminal nodes instead of a single result class. Decision trees are incrementally built by first selecting the test which splits the manually annotated training sample into the most homogeneous subsets with respect to the class.

Top-down induction of decision trees is defined recursively. At the beginning, we have a set of training examples and a set of possible features. The purpose of a decision tree is to order the examples and the algorithm should start with the best ordering features to reach a fast convergence. So a decision tree algorithm has its own feature subset selection. However, to find the best ordering features we have to rank all available features. Therefore, we can use an entropy-based measure like the information gain based on the input training set S and a single feature F, with following equation:

$$I_1 \quad (S, F) = E \quad (S) - A \quad e \cdot \quad (S, F).$$

The average entropy is defined by the following formula :

$$A \quad e \cdot \quad (S, F) = \sum_i \frac{|s_i|}{s} E \quad (S_i)$$

2.2 SVM Classifier

A SVM is a relatively new class of machine learning techniques that was first introduced by [8]. SVMs are a very popular technique for text categorization used in the machine learning community. They are considered to be one of the most effective classification methods according to their performance on text classification, as proven by many researchers [9,10]. Based on the structural risk minimization principle from computational learning theory, SVMs seek a decision surface to separate the training data points into two classes and to make decisions based on the support vectors that are selected as the only effective elements in the training set. Multiple variants of SVMs have been developed [3]. In this paper, our discussion is limited to linear SVMs due to their popularity and high performance in text categorization [11]. The optimization procedure of SVMs (dual form) is to minimize the following:

$$\vec{\alpha} = a \quad \left\{ -\sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \vec{x_i}, \vec{x_j} \rangle \right\}$$

$$S \cdot \quad : \sum_{i=1}^{n} \alpha_i y_i = 0; \quad 0 \le \alpha_i \le C$$

2.3 NAÏVE BAYES

naive Bayes technique is exhaustively used for text classification. Naive Bayes is the simplest form of Bayesian network, in which all attributes are independent given the value of the class variable. Given a table of feature vectors, the technique decides the rear possibility, where the term is related to multiple security levels, and assigns it to the security level with the maximum rear possibility. There are two used approaches: multi-nomial models and multi-variate Bernoulli models. Naïve Bayes is a stochastic model of generating documents makes use of Bayes' rule. To classify as the best class n* for a new document d, it computes:

Journal of Theoretical and Applied Information Technology
15th December 2016. Vol.94. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$p(s_{i,j}|d_i) = \frac{p(s_{i,j})p(s_{i,j}|d_i)}{p(d_i)}$$

2.4 K-Nearest Neighbor(Knn)

The K-nearest neighbor (KNN) is a typical example-based classifier that does not build an explicit, declarative representation of the document, but rely on the tags attached to the training documents similar to the test documents. As a result, KNN has been called lazy learners, since it defers the decision on how to generalize beyond the training data until each new query instance is encountered[12].Given a test document d, the system finds the K-nearest neighbors among training documents. The similarity score of each nearest neighbor's documents to the test documents is used as the weight of the security level in the neighbor's documents. If two documents have feature vectors with similar similarity scores, means that they may have the same security level. The weighted sum in KNN categorization can be written as in the following Equation

$$score(d, l_i) = \sum_{d_j=K\ (d)} sim(d, d_j)\,\delta(d_j, l_i)$$

Where KNN (d) indicates the set of K- nearest neighbors of document d, if d_jbBelongs to security level l_i, $\delta(d_j, l_i)$ equals 1, or other-wise 0. For test document d, it should belong to the security level that has the highest resulting weighted sum.

3. INTRODUCTION TO SECURITY LEVEL

For any organization or any firm or any bank there must be specific level of security applied on data and information of that organization, and no doubt it may be vary to different organizations depending upon their nature of work and distribution of work assignments of their employee [13]. In this section we are classifying to data into two following ways.

A) Policy of organization;

B) Nature/characteristics of data type.

There may be different work can be distributed depending upon the different policies of organizations and same way data and information can be divided or distributed depending upon their accessibilities and privacy or security purposes. The level of security can be decided based on the accessibility of data of authorization among the employee of the organization, which can be of five levels.

1) First or high level (00);
2) Second level (01);
3) Third level (10);
4) Fourth level (11);
5) Fifth or normal level.

Out of five levels first or high level is not normal level it means it is highly protected with other users to accessibility for security purposes of the organization. And level fifth is public level it means its open to accessible for everyone in the organization. (Like in OOPS programming data can be distributed among the user between public and private access specifier). And gradually from second to fourth level of security is for the order level of security which works between first to fifth level OR private to public level security.

Diagrammatically the organization of different level can be show in the following figure (1) below.

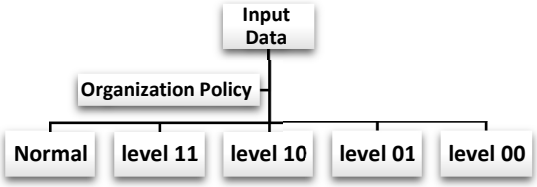

Figure 1 Security Level Classification

4. HYBRIDE CLOUD COMPUTING

A Hybrid Cloud incorporates the previous three types – public, private, and community clouds. It is of utmost importance in the case of Hybrid Cloud use that, while many resources are made available for the public, vital security issues are addressed by keeping important data and processes within a limited group, as shown in figure 2 below.

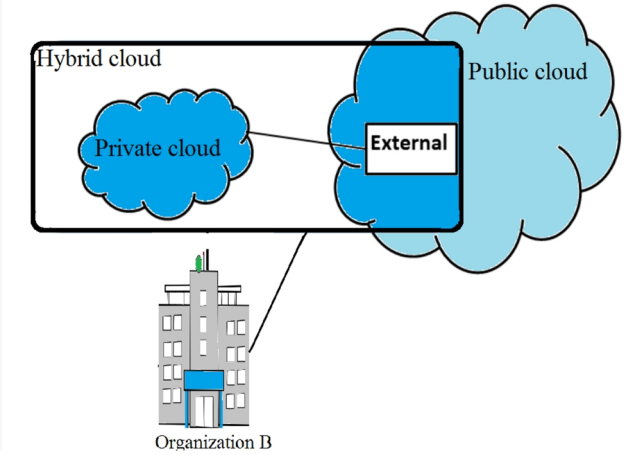

Figure 2 Hybrid Cloud

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. RESEARCH METHOD

Constructing an effective security level classification for confidential documents solution using a machine learning approach (classifiers combination using a Meta ensemble approach) requires many computational steps including data planning, preprocessing, feature selection and optimization, classification, and evaluation.

5.1 Preprocessing Phase

Using a supervised machine learning technique relies on the existence of annotated training data. Such data is usually created manually by humans or experts in the relevant field. The training data needs to be put in a format that is suitable to the solution of choice. New data to be classified also requires the same formatting. Depending on the needs of the solution, the textual data may need to be tokenized, normalized, scaled, and mapped to numeric classes, prior to being fed to a feature selection module. To reduce the training time with large training data, some techniques such as chunking or instance pruning (filtering) may need to be applied.

5.2 Feature Selection Methods

Feature selection method (FSS) is one of the most important tasks that will enhance the performance of sentiment classification system, as they will select the most predictive features. A FSS also reduces the high dimensionality of the data and removes irrelevant, redundant, and noisy data [15]. As consequence, FSS can help in building faster, cost effective and accurate security level classification models. In this work, selection of the FSSs algorithms is influenced by the data size, data consistency and the need to investigate the most efficient FSSs[16]. The performance of these feature selection algorithms will be explored and compared with three state-of-the-art machine learning algorithms. In this section, we provide a brief introduction to seven effective FSSs: IG and Chi-squared. These seven methods compute a score for each individual feature and then select a predefined size for the feature set as shown in figure 3 below.

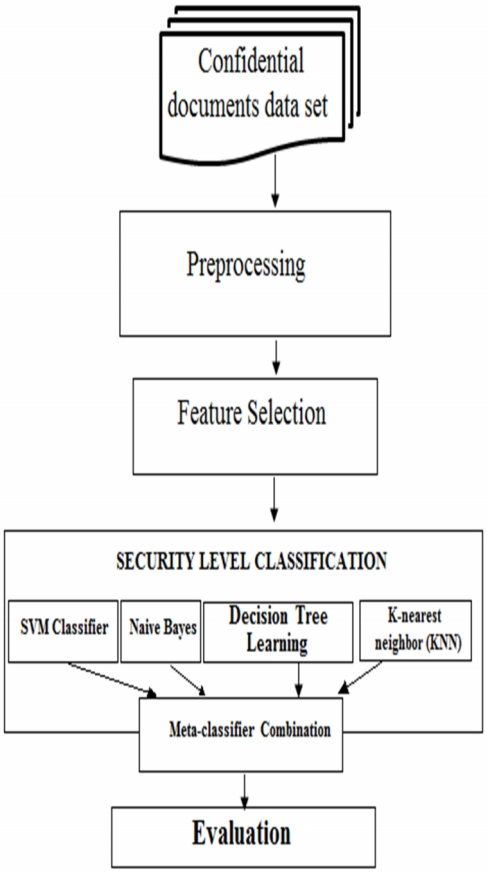

Figure 3 Feature Selection Methods

5.3 Collected Information

IG is a one of the most important feature selection algorithms for security level classification for confidential documents. It has been used as a term goodness measure in the field of machine learning [11]. IG is used to select important features with respect to class attribute. It is measured by the reduction in the uncertainty in identifying the class attribute when the value of the feature is known. The top ranked (important) features are selected for reducing the feature vector size in turn better classification results.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$IG(t) = -\sum_{i=1}^{|c|} p(c_i)\log p(c_i) + p(t)\sum_{i=1}^{|c|} p(c_i|t)\log p(c_i|t) + p(\bar{t})\sum_{i=1}^{|c|} p(c_i|\bar{t})\log p(c_i|\bar{t})$$

1) 3.1 Chi-squared statistic (χ^2)

The χ^2 statistic is one of the most commonly-used feature selection algorithms. The χ^2 statistic measures the lack of independence between the term and category [13] and is defined as follows:

$$\chi^2(c,t) = \frac{N \times (\text{AD-BC})}{(A+C)(B+C)(A+B)(C+D)}$$

$$\chi_{\max}{}^2(t) = \max_i(\chi^2(t,c_i))$$

where Ais the number of times that t and c co-occur,B is the number of times that t occurs without c,C is the number of times that c occurs without t, D is the number of times that neither c nor t occurs, and N is the total number of documents.

6. MACHINE LEARNING AND SECURITY LEVEL CLASSIFICATION FOR CONFIDENTIAL DOCUMENTS

As stated above, the two-layer ensemble of machine learning classifiers framework for handling the problem of security level classification is proposed. The two-layer framework consists of two stages of classification. In the first stage, Legal Documents Filtering based on ensemble of machine learning classifiers filters the documents to two categories legal and illegal. The second stage, a security level classification for confidential documents based also on ensemble of machine learning classifiers classifies legal documents into the five security levels. Figure 4 shows the architecture of the two-level ensemble classification Framework.

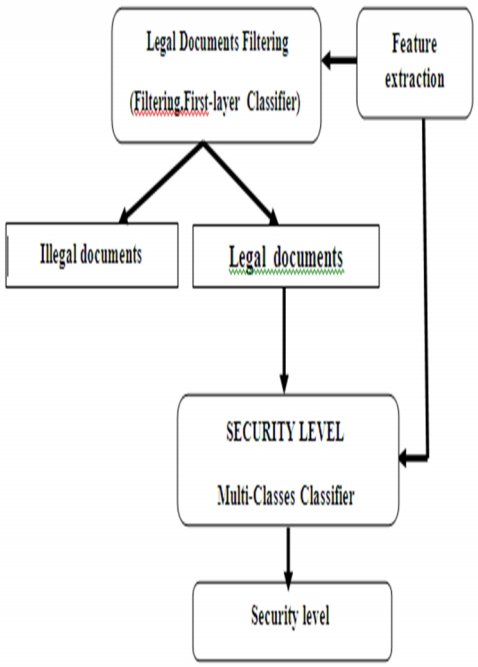

Figure 4 The Architecture Of The Two-Level Ensemble Classification Framework

In this paper, several experiments have been conducted in order to evaluate the proposed approaches. Several experiments have been performed to empirically evaluate two feature selection methods (Chi-square (χ2), information gain (IG)) and five classification methods (decision tree classifier, Support Vector Machine (SVM), Naïve Bayes (NB), and K-Nearest Neighbor (KNN) and meta-classifier combination) for Legal Documents Filtering. Table 1 show the results obtained. The results show that all classifiers perform better with the information gain feature selection methods than their results with Chi-Square feature selection method. Results also show that Support Vector Machine (SVM) outperforms achieve the best results among all individual classifiers. However, the proposed meta-classifiers method achieves the best results among all classification approaches. Almost all the machine learning based techniques and approaches have two phases, where the training is performed initially to produce a trained machine, and then a classification

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

step is performed. In this study, the following machine learning approaches are evaluated.

7. PERFORMANCE MEASURES

In this work, the performance measures are precision and recall. Let TP be the total number of documents correctly assigned to the security level by a system (True Positive). FP refers to the total number of documents incorrectly assigned to the security level by a system (False Positive). FN refers to the total number of documents ignored by the system but belong to the security level (False Negative). TN denotes the total number of documents correctly rejected by the system (True Negative). Precision and recall are evaluated for each security level.

Precision is a percentage measure of the capability of a system to retrieve only relevant items.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall is a percentage measure of the availability of all relevant items classified by the system.

$$Recall = \frac{TP}{TP + FN}$$

F-measure is the most popular measure for evaluating classification systems, where it combines precision and recall by function.

$$F_\beta = \frac{(1+\beta^2) * (precision * recall)}{(\beta^2 * precision + recall)}$$

8. RESULTS

Based on the determination of accuracy on different classification we are proposing that which one is the best and suitable method for applying to classification for security level of data. As a part of result following table is used to showing which are the best and suits on level of security. In the table if we see there are various security levels showing along with their percentage of security, but in case of "Meta-classifier combination" there is 92.64%,94.61%,93.895,94.58%,94.233% of Chi-Square and Information Gain respectively and that is the best and accurate in our observation and we proposed the same one. In table 1 as shown the performance of five classification methods (decision tree classifier, Support Vector Machine (SVM), Naïve Bayes (NB), and K-Nearest Neighbor (KNN) and meta-classifier combination) for Legal Documents Filtering.

Table 1 Classification Performance

Security Level	Chi-Square			Information Gain		
	P	R	F_β	P	R	F_β
Decision Tree Classifier	90.84	87.7	89.24239	89.99	93.03	91.48475
Support Vector Machine	92.92	90.66	91.77609	93.19	93.08	93.13497
Naïve Bayes	89.53	86.7	88.09228	88.02	93.33	90.59726
K-Nearest Neighbor	87.92	88.23	88.07473	89.99	93.03	91.48475
meta-classifier combination	92.64	94.61	93.61464	93.89	94.58	94.23374

In the second experiment, the decision tree ID3 classifier is applied on testing set using 10-fold cross-validation. This section shows the best results obtained when the decision tree ID3 classifier is applied. In table 2 shows the performance in terms of the precision, recall, F-measure of decision tree ID3 for of security level classification for confidential documents with the chi-square and information gain. Results show that work better with the Information Gain feature selection.

Table 2 the performance of decision tree ID3 for of security level classification for confidential documents with the chi-square and information gain .

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2 The Performance Of Decision Tree

Security Level	Chi-Square			Information Gain		
	80.56906	81.83177	81.19550605	P	R	F_β
Level 1	85.25845	84.04312	84.64642288	84.25845	83.04312	83.64637
Level 2	85.74067	87.05828	86.39445154	84.74067	86.05828	85.39439
Level 3	82.67896	84.8254	83.73842751	84.12927	82.88259	83.50128
Level 4	83.561785	84.4396425	83.99370199	82.76472	85.00764	83.87119
Average	80.56906	81.83177	81.19550605	83.97328	84.24791	84.10331

In the third experiment, the KNN classifier is applied on testing set using 10-fold cross-validation. This section shows the best results obtained when the KNN classifier is applied. Table3 shows the performance in terms of the precision, recall, F-measure of KNN for of security level classification for confidential documents with the chi-square and information gain. Results show that work better with the Information Gain feature selection. As noted from results reported on this experiment and previous experiment, the results obtained using KNN classifier is outperformed than that obtained using decision tree classifier. In table 3 as shown the performance of KNN for of security level classification for confidential documents with the chi-square and information gain:

Table 3 The Performance Of KNN

Security Level	Chi-Square			Information Gain		
	P	R	F_β	P	R	F_β
Level 1	79.1639	87.23813	83.00512	76.95977	90.78611	83.30312667
Level 2	85.28945	84.17344	84.72777	78.7448	91.80134	84.77328373
Level 3	78.30884	85.85454	81.90827	78.8188	90.23673	84.14218422
Level 4	79.20939	84.89808	81.95514	87.42833	82.37951	84.82886285
Average	80.4929	85.54105	82.89908	80.487925	88.8009225	84.26186437

In the fourth experiment, the SVM classifier is applied on testing set using 10-fold cross-validation. This section shows the best results obtained when the SVM classifier is applied. Table 4 shows the performance in terms of the precision, recall, F-measure of SVM for of security level classification for confidential documents with the chi-square and information gain. Results show that work better with the Information Gain feature selection... As noted from results reported on this experiment and two previous experiments, the results obtained using SVM classifier is outperformed than that obtained using all other individual classifiers. Results also show that Support Vector Machine (SVM) outperforms achieve the best results among all individual classifiers for security level classification for confidential documents. In Table 4 the performance of SVM for of security level classification for confidential documents with the chi-square and information gain

Table 4 The Performance Of SVM

Security Level	Chi-Square			Information Gain		
	P	R	F_β	P	R	F_β
Level 1	85.73175	85.77235	85.75204519	84.90257	86.76093	85.82169
Level 2	86.10545	84.89012	85.4934661	87.01532	86.01026	86.50987
Level 3	86.58767	87.90528	87.24150031	86.75711	87.90563	87.32759
Level 4	88.72833	83.67951	86.12999476	85.76472	88.00764	86.87171
Average	86.7883	85.561815	86.15425159	86.10993	87.17112	86.63272

In the fifth experiment, the NB classifier is applied on testing set using 10-fold cross-validation. This section shows the best results obtained when the NB classifier is applied. In table 5 shows the performance in terms of the precision, recall, F-measure of NB for of security level classification for confidential documents with the chi-square and information gain. Results show that work better with the Information Gain feature selection. As noted from results reported on this experiment and three previous experiments, the results obtained using NB classifier is worse than that obtained using all other individual classifiers. Results also show that NB achieves the worst results among all individual classifiers for security level classification

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

for confidential documents. In table 5 as shows the performance of NB for of security level classification for confidential documents with the chi-square and information gain:

Table 5 The Performance Of NB

Security Level	Chi-Square			Information Gain		
	P	R	F_β	P	R	F_β
Level 1	73.91429	87.28007	80.04305368	76.42859	90.9401	83.0552431
Level 2	78.02424	85.72585	81.69393122	78.39054	86.26514	82.13954001
Level 3	76.84181	91.979	83.73177266	76.56207	87.4305	81.63613829
Level 4	87.72833	82.67951	85.12912713	87.72833	82.67951	85.12912713
Average	79.1271675	86.9161075	82.64947117	79.7773825	86.8288125	82.99001213

Finally, the sixth experiment has been performed to empirically evaluate the proposed meta-classifiers combination method which combines the four classifiers (decision tree classifier, Support Vector Machine (SVM), Naïve Bayes (NB), and K-Nearest Neighbor (KNN)) for security level classification for confidential documents. The meta-classifier combination method is applied on the testing data set by using 10-fold cross-validation. Table 6 shows the performance in terms of the precision, recall, F-measure of meta-classifiers combination method for of security level classification for confidential documents with the chi-square and information gain. Results show that meta-classifiers combination method work better with the Chi-Square feature selection. As noted from results reported on this experiment and three previous experiments, the results obtained using meta-classifier combination method is outperformed that obtained using all individual classifiers (decision tree classifier, Support Vector Machine (SVM), Naïve Bayes (NB), and K-Nearest Neighbor (KNN)) for security level classification for confidential documents.

10. Conclusion

Cloud computing is suffering from various problems which impacted the security level of cloud computing in two phases. First, the organizations or institutions are very concerned in improving the security of cloud computing through the application of the authority model and dynamic classification of data model based on the multi-level security. Second, they prefer to develop the multi-key cipher algorithm to manage the encryption based on the level of security. Most of organization must apply new policies in classifying the data into many security levels based on the nature of data to save time, and effort [1]. However, the results obtained using meta-classifiers combination methods are significantly higher than that obtained using the best individual SVM classifier. These results indicate that the proposed method meta classifier combination which combine four classification methods are most suitable technique for security level classification. In table 6 the performance of meta-classification for of security level classification for confidential documents with the chi-square and information gain. The limitations of this study included environment validation which impacted on the accuracy of classification. In addition, image, video, and audio hadn't been covered when the test was administered.

Table 6 The Performance Of Meta-Classification

Security Level	Chi-Square			Information Gain		
	P	R	F_β	P	R	F_β
Level 1	89.49452	87.8778	88.67879	89.92886	88.13754	89.02419
Level 2	90.46544	89.10433	89.77973	93.1738	88.49685	90.77512
Level 3	92.7346	88.89353	90.77345	90.16025	86.67404	88.38278
Level 4	91.00203	89.88213	90.43861	93.22283	91.2038	92.20226
Average	90.92415	88.93945	89.91765	91.62144	88.62806	90.09609

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFERENCE

[1] Mohanaad Shakir, Asmidar Abubakar, Younus Yousoff, Ali Makki Sagher, Hussam Alkialy," Diagnosis Security Problems for Hybrid Cloud Computing in Business cloud", Journal of Theoretical and Applied Information Technology, E-ISSN: 1817-3195,ISSN: 1992-8645,. Vol.90. No.2, 31 Aug. 2016, P 151-157

[2] Erdem ALPARSAN,"*SECURITY LEVEL CLASSIFICATION FOR CONFIDENTIAL DOCUMENTS BY USING ADAPTIVE NEURO-FUZZY INFERENCE SYSTEMS*", The Graduate School of Natural and Applied Sciences Computer Engineering(Bahçeşehir Üniversitesi), Master's Thesis, 2010

[3] Joachims T., 1998. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. European Conference on Machine Learning

[4] Feldman R. & Sanger J., "*Text mining handbook"*. Cambridge: Cambridge University Press, 2007.

[5] Han J.W. & Kamber M."*Data mining concept and techniques"*. Second Edition. San Francisco: Elsevier, 2007.

[6] L. Breiman, J.H. Friedman, R.A. Olshen, & P.J. Stone, "*Classification and Regression Trees*", Wadsworth, Belmont, CA, 1984.

[7] J. R. Quinlan, "*C4.5: Programs for Machine Learning"*, Morgan Kaufmann, San Mateo, CA, 1993.

[8] C. Cortes, V. Vapnik," Support-Vector Networks", Kluwer Academic Publishers, Boston. Manufactured in The Netherlands, 20,273-297(1995).

[9] Isa, Dino and Hong, Lee Lam and Kallimani, V.P. and Rajkumar, R. "*Text document pre-pro*cessing *using the Bayes formula for classification based on the vector space model*". Computer and Information Science,1 (4). pp. 79-90. ISSN 1913-8989,2008

[10] T. Joachims," *A Statistical Learning Model of Text Classification forSupport Vector Machines*", GMD Forschungszentrum IT, AIS.KD Schloss Birlinghoven, 53754 Sankt Augustin, Germany,2001

[11] Y. Yang and J. Pedersen."*A comparative study on feature selection in text categorization*". In Proceedings of ICML-97, the 14th International Conference on Machine Learning, 1997.

[12] Y. K. Lee, H. T. Ng, T. K. Chia ,*" An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation"*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) , pages 41–48,2002.

[13] R. Shaikh , M. Sasikumar," *Data Classification for achieving Security in cloud computing*", Procedia Computer Science 45 493 – 498, 2015.

[14] L.i Galavotti, F. Sebastiani, M.Simi," *Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization*", Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2000),page 59-68, Springer,2000

[15] CAROPRESO, M. F., MATWIN, S., AND SEBASTIANI, F.. A *"learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In Text Databases and Document Management"*: Theory and Practice, A. G. Chin, ed. Idea Group Publishing, Hershey, PA, 78–102, 2001.

[16] M. Dash, H. Liu*," Feature Selection for Classification*", Intelligent Data Analysis 1 131–156, 1997.