GROUP IDENTIFICATION BASED CLASSIFICATION
TECHNIQUE FOR AGGREGATED COMMON DATA TO SAVE MEMORY SPACE IN DIGITAL FORENSIC INVESTIGATION (DFI)

Amit Pratap Singh¹ and Dr. R. C. Jain²
¹Research Scholar, ²Director, Department of Computer Applications, Samrat Ashok Technical Institute, Vidisha, M.P., India

ABSTRACT
The problem of assigning pre-defined class labels to incoming, unclassified data is called as classification. This is the domain which could be well used in forensic digital investigation. Sample of pre-classified data is the base for the class labels. There are a number of approaches that have been proposed for classification like probabilistic, information retrieval based and machine learning. And these classification methods play very vital role for forensic digital investigation. This data within a group bind to a set of patterns, and that these patterns closely correspond to, and are derived from the data of the particular group which would be very useful while handling forensic digital investigation. Proposed work shows that the effectiveness of the proposed forensic digital investigation method. This forensic digital investigation based on classification method produces the same results of forensic digital investigation while improving the memory utilization. The proposed work is implemented in MATLAB.

Keywords
Classification, Memory Utilization, Weblog, XML, DFI.

[1] INTRODUCTION
The task of classification occurs in a wide range of human activity. At its broadest, the term could cover any context in which some decision or forecast is made on the basis of currently available information, and a classification procedure is then some formal method for repeatedly making such judgments in new situations. The construction of a classification procedure from a set of data for which the true classes are known has also been variously termed pattern recognition, discrimination, or supervised learning (in order to distinguish it from unsupervised learning or clustering in which the classes are inferred from the data).

[2] CLASSIFICATION
Classification has two distinct meanings. We may be given a set of observations with the aim of establishing the existence of classes or clusters in the data. Or we may know for certain that there are so many classes, and the aim is to establish a rule whereby we can classify a new observation into one of the existing classes. The former type is known as Unsupervised Learning (or Clustering), the latter as Supervised Learning. The supervised methods used are Naïve Bayes classification, J48 Decision Trees, and SVM, from the class of unsupervised learning a technique of adaptation of k-clustering.
[2.1] Type of classification

[2.1.1] Naïve Bayes Classifier

The Naïve Bayes classifier follows the simple principle and innate idea. In various situations it has been found that the Naïve Bayes, operate in a better way than any other proposed complex algorithm. This method utilizes the variable present in data sample by monitoring in an isolated and independent manner.

The Naïve Bayes depends on the principle of Bayes of conditional probability. This effectively uses the entire attribute comprises by the data, and operate each variable independent from one another as each has its own importance of existence. As instance, taken into consideration a set of training data which posses different species of animal (as monkey, zebra and elephant), and when any new species or new instances encounter, then based on this available data attributes we have to classify the new instance. Now we know that the attributes of animal are having huge size, small tail, long trunk and others. Whereas Monkeys attributes are smaller in size, jump, and can climb the trees, zebra comprises the strip of black and white throughout the body.

The Naïve Bayes classifier observes all the available attributes independently to classify the new encountered instance. Thus if new instance belongs to the class of monkey, then Naïve Bayes classifier will not classify based on whether it can climb trees, small in size, or jump. However it verifies each attribute independently that new instance can climb trees, and it jump, and is small in size. This simple follow the principle that the function of one attribute is entirely different from the other attribute. From the outcomes of various experiments it has proved that the performance of Naïve Bayes is equitable with other method of classification. Thus this classification method has similar accuracy in many cases as those obtained by the support vector machine and J48 decision tree algorithm.

[2.1.2] J48 Decision tree

The J48 decision tree [1, 2] a kind of machine learning method based on prediction which take the decision of the target value of novel sample on the basis of attributes of available data. The tree structure and its terminology represent different meaning, such as internal node of tree indicate the different attributes, the edge between the node represent the possible values which an attribute retain from observed sample, the leaf node of the tree represent the outcome value of the dependent variable.

The dependent variable is the variable, whose value is forecasted, thus by name clearly understand that its value is derived (decided or depend) from the value of other attribute. The attribute called as the independent variable which helps in deriving the value of the dependent variable.

The step conducted during the J48 classification decision tree are as follow: For obtaining the classification of any given novel item, then the first thing is to construct the decision tree on the basis of the available data value of the different attributes. By this when any new item set also called as training data is encounter then it recognized the attribute that discriminate the instances preciously and clearly. This helps us to have more knowledge about the data instance, thus make it easier to classify the data in best possible manner. Then, from all possible values of identified feature; if there exist any values which does not comprises any ambiguity, or in other word the new data instance is belonging within its class or category then have similar value as the target variable, then we simply end the branch of classification and assign the obtained value to leaf node of the tree.

[2.1.3] Support Vector Machines
Another well-known method of classification is Support Vector Machine [2], also used for regression. One of the major benefits of using this method is that it utilizes some kernel that easily transforms the given problem, for example the linear method of classification is applied over nonlinear data. Thus by using kernel equation this technique organize the data instances in within the multidimensional space, and the present hyper plane separate the data instances of one type from the other.

To transform the linearly non separable data present in one domain into other uses the kernel equation which is nothing but a transformation function. This transformation function is linear, Gaussian, or something which helps us in obtaining our condition. First thing which we have to perform is to divide the data into two different categories, then next to have such hyper plane which effectively separate the two instances. Here the role of this hyper plane is important as this is responsible for deciding the target value for future forecast. The hyper plane is chosen so it maximizes the margin on both side of the support vector present. The instances which are present on either side of the hyper plane or a few on wrong side is referred as the support vector. The diagram represent in the figure 1 depict a clear picture of the support vector and hyper plane.

![Figure 1: Support vector machine](image)

**[3] DIGITAL FORENSIC INVESTIGATION**

Today’s era is the world of computer, as with the rapid growth of computer the usage of it also increases with the similar rate. Thus, all sector of organization whether governmental or private are dependent on computer and Internet for ease of their work and take their business to crest. Within the past few years, a new class of crime scenes has become more prevalent, that is, crimes committed within electronic or digital domains, particularly within cyberspace. Criminal justice agencies throughout the world are being confronted with an increased need to investigate crimes perpetrated partially or entirely over the Internet or other electronic media. Resources and procedures are needed to effectively search for, locate, and preserve all types of electronic evidence. This evidence ranges from images of child pornography to encrypted data used to further a variety of criminal activities. Even in investigations that are not primarily electronic in nature, at some point in the investigation computer files or data may be discovered and further analysis required.

In this modern age, it is hard to imagine a crime that does not have a digital dimension. Criminals, violent and white-collar alike, are using technology to facilitate their offenses and avoid apprehension, creating new challenges for attorneys, judges, law enforcement agents, forensic examiners, and corporate security professionals. Terrorists are using the Internet to communicate, recruit, launder money, commit credit card theft, solicit donations, and post
GROUP IDENTIFICATION BASED CLASSIFICATION TECHNIQUE FOR AGGREGATED COMMON DATA TO SAVE MEMORY SPACE IN DIGITAL FORENSIC INVESTIGATION (DFI)

propaganda and training materials. Computers played a role in the planning and subsequent investigations of both World Trade Center bombings. Ramsey Yousef’s laptop contained plans for the first bombing and, during the investigation into Zacarias Moussaoui’s [CASE EXAMPLE (MASSACHUSETTS, 2005–2010)] [17] role in the second attack, over 100 hard drives were examined. Islamist extremists are going so far as to develop their own tools to avoid detection and apprehension, including a program named “Mujahideen Secrets 2” [18] designed to encrypt e-mail and Instant Messaging communications. Their use of the Internet creates challenges for digital investigators and requires more international legal cooperation and information sharing.

[4] RELATED WORK

Sarac, E and Ozel,S.A[3]: In this paper authors present firefly Algorithm brought into existence by Xin-She Yang in 2007-2008 at Cambridge University, this method has coined its idea from the nature and behavior of fireflies. The assumption introduced in this method are:

As fireflies are unisexual so they an attraction to each other despite of their sex behavior.

This attractive feature of fireflies is directly dependable on their brightness and these both factor have impact of distance on it with an increase in the distance they decrease. Since if there are two fireflies with different brightness, then the less brighter fireflies will move closer to more brighter, in case if brightness are equal then there movement are random.

We evaluate the brightness by the measure of the objective function, and we define a the attraction measure $\beta$ with a distance factor $r$ as

$$\beta = \beta_0 e^{-\gamma r^2}$$

The motion of firefly $I$ which get attracted by the brighter firefly $j$ is given by the term as follows:

$$X_{it+1} = X_{it} + \beta_0 e^{-\gamma r^2} (X_{it} - X_{jt}) + \alpha t \epsilon_{it}$$

Another factor indicates an attraction. The third factor is randomization $\alpha t$ being the randomization parameter, and $\epsilon_{it}$ indicate a random number vector evaluate by a Gaussian or uniform dispersion at an interval of time $t$. In case if $\beta_0 = 0$, it considers a usual tour. On other part, if $\gamma = 0$, it curtails to an alternative of particle swarm optimization.

Aliakbary, S. Abolhassani, H., Rahmani, H. and Nobakht, B.[4]: Social tagging is a process in which many users add metadata to a shared content. Through the past few years, the popularity of social tagging has grown on the Web. In this method we investigated the use of social tags for Web page classification: adding new Web pages to an existing Web directory. A Web directory is a general human-edited directory of Web pages. It classifies a collection of pages into a wide range of hierarchical categories. The problem with manual construction and maintenance of Web directories is the significant need of time and effort by human experts. Our proposed method is based on applying different automatic approaches of using social tags for extending Web directories with new URLs.[5]

Gowri. R and Lavanya, R. [5]: In this paper author describe a brief survey about the existing approaches in web services composition. The main research areas in web services are related to discovery, security, and composition. Among all these areas, web services composition turns out to be a challenging one, because within the service-oriented computing domain, Web service composition is an effective recognition to satisfy the hastily changing requirements of business. Therefore, the Web service composition has unfolded broadly in the research side. However, the current attempts to classify Web service composition is not appropriate to the objectives. This paper proposes a novel classification matrix for Web service composition that distinguishes between the context and technology dimension. The context dimension is aimed at analyzing the QoS influence on the effort of Web service composition, while the technology dimension focuses on the technique influence on the effort. Finally, this paper provides a suggestion to improve the
quality of service selection those participates in the composition process with Cskyline approach using agents. [6]

Jinbeom Kang and Joongmin Choi [6]: This paper present various analysis are actively study on mining web data from the numerous available data on WWW. As the web page are not fully structures so it become difficult to determine from it’s the informative block, methods which provide the useful data extraction from the useless data such as advertisements is more important. Commonly a web page has many different blocks in which it include data and structural facts. In this proposed method we introduce a web page classification in form of blocks by constructing a Tree Alignment model that indicate the HTML feature and a vector model that represents a feature of blocks. The different websites have their personal templates and blocks may be linked by category nevertheless they are placed at similar position in the web browser or have a similarity in their structure. Thus, by constructing the single classifier it becomes difficult to classify a block accurately. To overcome this problem in our proposed method we use the multiple classifier one for each training data set and classification method succeed by combining all of them. [7]

Kovacevic, M. Diligenti, M. Gori and M. Milutinovic, V. [7]: In this paper author have studied that with the fast development of Internet, the Web has become the largest information source for people. Searching information through Internet becomes a more and more popular activity. However, useful information is often accompanied by a large amount of noises. Almost all web pages on the Internet contain noises irrelevant to the main content, such as navigation bar, copyright information, survey or feedback questionnaire etc. These noises affect the efficiency of algorithms for web page classification, clustering, information extraction and searching although they could be useful for other purposes, such as to ease browsing the web pages. It is important to distinguish the informative blocks from the noisy blocks. In this method, we use block to denote the semantic part of the web page, and informative block to denote the main content. Web page information extraction aims to identify and extract relevant data from web pages and put the extracted data into some particular format so that the extracted information can be easily used by some software applications such as Search Engine

[5] FORENSIC INVESTIGATION PROCESS

The entire process of digital forensic comprises 4 major phases:
Collection: In this phase of digital forensic we recognize, label, and aggregate the data from all possible sources and also take an action that preserve the integrity of data collected.
Examination: This phase uses the manual and automated method on aggregated data to process it forensically, thus retrieve the data that is of particular interest, while preserving the integrity of it.
Analysis: This phase analyzes the ramification obtained from the examination phase by using the justifiable methods, how impetus the derived information to answer the question is depends on the collection and examination phase.
Reporting: Last phase is report the outcome of analysis such as illustrate the used action, describe the usage of tools and procedure adopted, evaluate the other action requires to be executed (e.g., examination of other available data source, protect the recognized vulnerabilities, enhance the existing security measure), and confront the recommendation should be adopted to enhance the tools, procedure, rule of forensic process.
GROUP IDENTIFICATION BASED CLASSIFICATION TECHNIQUE FOR AGGREGATED COMMON DATA TO SAVE MEMORY SPACE IN DIGITAL FORENSIC INVESTIGATION (DFI)

[6] AGGREGATED COMMON DATA (ACDACDGIC-FDI) BASED GROUP IDENTIFICATION BASED CLASSIFICATION FOR DIGITAL FORENSIC INVESTIGATION

In this session, it provides the details and description of the proposed method. It illustrates the use of the Proposed Work (ACDACDGIC-FDI) and how it works for the datasets. To maintain the Forensic Digital Investigation Through classification result while reducing the memory required is the goal of this work. The methodology for evaluating the above objective of the proposed technique is also discussed in this session. The proposed method of classification is discussed in the algorithm mentioned below (fig 2).

Fig 2: Proposed Classification Method

[7] RESULT AND ANALYSIS
The system used for execution of the proposed method for forensic digital investigation by classification is as follows:

<table>
<thead>
<tr>
<th>Model</th>
<th>Pentium i3 CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAM</td>
<td>4 GB</td>
</tr>
<tr>
<td>64 Bit Operating System</td>
<td></td>
</tr>
<tr>
<td>Windows 8</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 1: System Configuration

In order to evaluate the classification and its efficiency of the various data records through this propose method, two web log datasets are taken into consideration.

NASA weblog file[8]
comdotzone weblog file [9]

The input common XML file contains total 90860 records which gets classified by proposed method. The structure of the CommonFormat file is shown in figure 3.
Figure 3 show the structure of input record set ‘CommonFormat’. The output file ‘agCommonFormat’ contains total only 3 records after proposed method. Figure 4 show the structure with records of output record set. Here the aggregated file ‘agCommonFormat’ by Date (expanded nodes) is shown.

Figure 4 show the structure with records of output record set ‘agCommonFormat’ aggregated file by Date (expanded nodes).

Aggregated file ‘agCommonFormat’ by Date (collapsed nodes) – having 3 lines of 3 dates and it is shown in figure 5.

**TABLE 2: Forensic Digital Investigation by Classification result of ‘agCommonFormat’ and ‘agCommonFormat’ files**

<table>
<thead>
<tr>
<th>Data Size</th>
<th>Common Input File</th>
<th>Aggregated Input File</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>59.946</td>
<td>59.946</td>
</tr>
<tr>
<td>Half</td>
<td>78.80</td>
<td>78.80</td>
</tr>
</tbody>
</table>
Aggregated file ‘agCommonFormat’ and ‘CommonFormat’ takes memory in system which is shown in TABLE II and Graph 2.

Graph 1: Show the comparative classification results before and after aggregation method on common file.

<table>
<thead>
<tr>
<th>Common Input File</th>
<th>Aggregated Input File</th>
</tr>
</thead>
<tbody>
<tr>
<td>5099</td>
<td>1976</td>
</tr>
</tbody>
</table>

TABLE 3: Memory Utilization of ‘agCommonFormat’ and ‘agCommonFormat’ files

Graph 2: Show the comparative memory utilization results before and after aggregation method on common file.

[8] CONCLUSION
This work investigates that forensic digital investigation through classification and memory utilization problem. From the session 6, researchers have shown and proved that the proposed algorithm works with great efficiency over these parameters for the taken dataset. From the graph 1 and 2 it is very clear that even after aggregation this method is capable enough to maintain the forensic digital investigation results while aggregated common file save memory by 60% (almost).

REFERENCES
[1] Arash Rezaei, and Behrouz Minaei-Bidgoli, “Comparison of Classification Methods Based on the Type of Attributes and Sample Size”, Department of Computer Engineering, Iran University of Science & Technology (IUST), Tehran, Iran.


Author[s]
Mr. Amit Pratap Singh  is a Research Scholar  in Barkatullah University, Bhopal, Madhya Pradesh, India. His research center is, Samrat Ashok Technological Institute, Vidisha. He is doing research in the area of Web Usage Mining, Cyber Crime.

Dr. R. C. Jain Director (Retd.), Samrat Ashok Technological Institute, Vidisha. He has 30 years teaching experience and 15 years research experience. He has presented research papers in more than 100 national and international Conferences and published more than 100 papers in national and international journals. His research areas include Data mining, web mining and Network security.