# Mixed-sampling Approach to Unbalanced Data Distributions: A Case Study involving Leukemia's Document Profiling

Wu QingQiang
School of Software
Xiamen University
Xiamen, Fujian Province, P. R. China, 361005
wuqq@xmu.edu.cn

Liu Hua
Information resource center
Institute of Scientific and Technical Information of China
Beijing, P. R. China, 100038
liuhua@mail.las.ac.cn

Liu KunHong* (corresponding author)
School of Software
Xiamen University
Xiamen, Fujian Province, P. R. China, 361005
lkhqz@xmu.edu.cn

*Abstract:* - Leukemia's types and their relationships to literatures are introduced, based on which data set about Leukemia for classification is constructed with original data sources, such as Cancer Gene Census, PubMed and gene2pubmed. The data set is imbalanced as the research object. Based on the introduction of current classification methods of imbalanced data set, the problems of sampling in imbalanced data set are analyzed, and mixed-sampling method is proposed to classify the Leukemia data set. The multi-class problem about Leukemia is transferred to a set of two-class problems. Area Under Receiver Operating Characteristic (ROC) Curve (AUC) are used to evaluate the mixed-sampling method. Then, experiments are performed to verify the classification efficiency and stability of eight classification methods, and their classification results are comparatively analyzed. It can be found that the mixed-sampling method achieves the best performance. At last, the research work in this paper is concluded with a look forward to the future work.

*Key-Words:* - Leukemia, Literature Profiling, Imbalanced Data Distribution, Decision Tree, mixed-sampling, Ensemble Learning.

## 1 Introduction

Leukemia was first identified by Germany pathologist Rudolf Virchow in 1847. It is a type of cancer that affects the blood and bone marrow, the spongy center of bones where our blood cells are formed. The disease develops when blood cells produced in the bone marrow grow out of control. Currently many types of Leukemia have been discovered, of which the most common types are: Acute Myeloid Leukemia (AML)[1], Acute Lymphoblastic Leukemia (ALL)[2], Chronic Myeloid Leukemia (CML)[3] and Chronic Lymphocytic Leukemia (CLL)[4]. Each main type of leukemia is named according to the type of cell affected (a myeloid cell or a lymphoid cell) and whether the disease begins in mature or immature cells. Other types of leukemia and related disorders include: Hairy Cell Leukemia (HCL), Chronic MyeloMonocytic Leukemia (CMML) and Juvenile MyeloMonocytic Leukemia (JMML)[5].

Many facts show that the Leukemia is caused by the abnormal genome structures or functions, just like other cancers. Cancer genes are confirmed to be related to Leukemia, including c-myc[6], MDM2[7], c-fos[8], BCL-1, BCL-2 and BCL-3[9], and P53[10].

Researches on Leukemia and related genes are published on the literatures which include considerable information of Leumekia, genes and their relationships[11].

These literatures play an important role in helping humans understand and treat the Leukemia. However, there are so many of them published each year that researchers can hardly obtain useful information. So emerges the need for automatic classification to these literatures. In this paper, we try to classify the related literatures by the way of constructing classifiers with document classification data related to leukemia. Thus, researchers may use this classifier to classify all the Leukemia literatures automatically according to leukemia types, and when new literatures are available, they can be classified into certain Leukemia's type, thus researchers can get the very part of literature related to their research.

As the document data combine to form an unbalanced multi-class data set, we find that the classification task is hard to solve using traditional classifier. As it has been proved that a classifier ensemble system is more robust than an excellent single classifier in many fields, many researchers designed different ensemble systems to deal with different problems successfully [16, 46-47, 52-53]. So besides the mechanism for tackling data imbalance problem, we also apply different ensemble learning methods to further improve the final classification accuracy.

There are six sections in this paper. Section one introduces the types of Leukemia, literatures about Leukemia genes, genes related to Leukemia, their relationships and the importance to the classification of literatures about Leukemia. Section two introduces the research object and elaborates the process for acquiring related data from Cancer Gene Census, PubMed and gene2pubmed, and construct classification data set of Leukemia with these three data sources. Section three introduces current classification methods of imbalanced data set. Section four proposes the mixed-sampling classification method based on analysis of problems about sampling the imbalanced data set. The procedures, related technologies and evaluation indicators are described in detail too. Section five is the section about experiments for verification. Eight classification methods, including mixed-sampling, are used to classify the imbalanced data set. And the results are compared and analyzed to verify the classification result and stability of mixed-sampling. The conclusions and future works are demonstrated in section six.

# 2 Research Object

The literatures about Leukemia's gene are selected as the research object. The classifier is constructed with mixed-sampling, and then used to classify the literatures. The process of aquiring original data from Cancer Gene Census, PubMed and gene2pubmed and constructing data set of gene literature about Leukemia, is described as follows.

## 2.1 Cancer Gene Census

The Cancer Gene Census (CGC)[17] is an ongoing effort to catalogue those genes for which mutations have been causally implicated in cancer by Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/). It's a public resource and can be downloaded freely from the website (http://www.sanger.ac.uk/genetics/CGP/Census/).
CGC contains those genes from the literature for which mutations have been causally implicated in cancer. Only the Mutated genes causally implicated in human cancer are selected into the list of CGC, that is, the relationships of gene and disease in the CGC list are sufficient evidence, true and reliable[18].

The samples (only the items related to this paper) of CGC are listed in Table 1.

CGC is well known for the scientists and researchers. The genes that have been selected for curation are taken from the list of cancer genes assembled in the Cancer Gene Census[19]. CGC also is used to generate a map of human cancer signaling[20] and to identify the novel cancer gene based on the gene network [21].

## 2.2 PubMed Literatures

PubMed is the U.S. National Library of Medicine's premiere search system for health information. PubMed comprises more than 20 million citations for biomedical literature from MEDLINE, life science journals, and online books. PubMed citations and abstracts include the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and preclinical sciences. PubMed also provides access to additional relevant Web sites and links to the other NCBI molecular biology resources (http://www.ncbi.nlm.nih.gov/pubmed/).

PubMed includes almost all the literatures in medicine, and it is used as research objects to analyze research topics within medical field by

many researchers[22]. Each literature in PubMed is identified by a unique value of PMID field, with which the data about literatures can be downloaded, such as TI, AU, AB and MH etc.

## 2.3 Gene2pubmed

Gene2pubmed (ftp://ftp.ncbi.nih.gov/gene/DATA/) stores the gene ID and its relationship with PubMed literature ID, and it is often used to obtain literatures associated with a particular gene.

Thomas, P. etc. introduced an evaluation strategy by using the NCBI Gene2Pubmed mapping as gold-standard[23]. Xu, H., et al. used the gene2pubmed file to extracted the articles related to the genes [24]. The gene2pubmed file is ASCII text file, its' format is showed as Table 2.

As shown in Table 2, the file of gene2pubmed contains three fields: taxonomy identifier, gene identifier and PubMed literatures identifier, which represents taxonomy ID in gene taxonomy, gene ID and PubMed ID in literature related to that gene in respective. Only gene ID and PubMed ID are used in this paper, whose characteristics are listed as follows:

- ✓ A gene ID might correspond to many PubMed IDs. For example, gene ID 1343045 corresponds to both PubMed IDs 9593780 and 10678977, as shown in Table 2. In other words, one gene might be studied in many literatures.
- ✓ Many gene IDs might correspond to the same PubMed IDs. For example, the PubMed ID 9873079 correspond to both gene IDs 1246502 and 1246505, as shown in Table 2. In other words, many genes might be studied in the same literature.

Therefore, the relationship between gene ID and PubMed ID is many to many (m:n).

## 2.4 The Process of constructing data set

The process of constructing Leukemia's gene literature data set is shown in Figure 1. The steps are described as follows:

Step 1: The gene IDs related to Leukemia are extracted from CGC file, which form a table with Leukemia's type named Leukemia2GeneIDRel.

Step 2: A list of gene IDs named GeneIDList is obtained from Leukemia2GeneIDRel. And all the PubMed IDs corresponding to GeneIDList are extracted from gene2pubmed index file. Then a table listing relationships between gene IDs and PubMed IDs is formed and named Gene2PubMedRel.

Step 3: A list of PubMed IDs can be extracted from Gene2PubMedRel, which is named PubMedIDList. The literatures about Leukemia can be downloaded from PubMed website based on the PubMed IDs in the PubMedIDList. All those literatures downloaded form a data set about Leukemia called LeukemiaSet.

Step 4: To analyze the literatures in LeukemiaSet. Four common Leukemia's types and their combinations are selected as the Leukemia's classification target based on the data in PubMed2LeukemiaRel and Gene2PubMedRel, and the distribution of Leukemia's literatures. And then all the literatures about the 6 categories are selected out to be the research object. The 6 categories are AML, ALL, CML, CLL, AML+ALL and CML+CLL[25] [26] [1]. Other types of Leukemia's are not included for there is few corresponding literatures.

Therefore, the classification data set about Leukemia are obtained, which are the basis of further research in this article.

## 2.5 Experience of Creating Dataset

The creation of Leukemia data set should be done based on the processes as shown in Figure 1. And the detail steps to create that data set in practice are listed as below.

Step 1: The CGC file was downloaded from http://www.sanger.ac.uk/genetics/CGP/Census/ on March 27, 2011, named Table_1_full_2011-03-22.xls, indicating the file is updated on March 22, 2011.

Step 2: Relationships between the Leukemia's types and gene IDs is obtained from Table_1_full_2011-03-22.xls, and stored in Leukemia2GeneIDRel.

1 Leukemia's type including AML, ALL, CML and CLL can be diagnosed by using gene expression profiling[20]. But there is intersection among four basic types of Leukemia such as AML, ALL, CML and CLL. For example: Acute Myeloid or Lymphoid Leukemia (AML+ALL) and Chronic Myeloid or Lymphoid Leukemia (CML+CLL) are adult leukemia, which can also treated as separate Leukemia's types[21]. At the same time, such categorization can avoid fuzzy classification in classifying the data set. Therefore, AML+ALL and CML+CLL are treated as separate type of Leukemia.

Step 3: A list of gene IDs related to Leukemia is obtained from Leukemia2GeneIDRel and named GeneIDList. There are 117 genes in GeneIDList, such as ABL1, ABL2, AF15Q14, AF1Q and AF3p21 etc.

Step 4: The file is downloaded freely from ftp://ftp.ncbi.nih.gov/gene/DATA/ on March 27, 2011 and named gene2pubmed.gz.

Step 5: All the PubMed IDs related to the gene IDs in GeneIDList are retrieved from gene2pubmed.gz. Then a table is created listing the relationships between gene IDs and PubMed IDS, which is named Gene2PubMedRel.

Step 6: A list of PubMed ID is created from Gene2PubMedRel and named PubMedIDList, which contains 22232 PubMed IDs.

Step 7: The literatures identified by the 2232 PubMed IDs in PubMedIDList are downloaded from PubMed website, thus forming the data set about Leukemia's gene literatures called LeukemiaSet.

Step 8: After initial analysis of LeukemiaSet, 2093 literatures about AML、ALL、CML、CLL and their combinations are selected as the research object. And 139 literatures about other Leukemia's types are discarded, for the research topics involved are distributed too broadly.

After all the steps mentioned above, the Leukemia's literature data set is obtained and its' distribution is shown in Table 3.

As we can see from Table 3, there are significant differences among the number of Leukemia literatures of different types. Of the six types, maximum number of literatures is 1369, which is related to Leukemia's Type 'ALL', while the minimum number is 12, which is related to Leukemia's Type ' CML+CLL'. The former one is 100 times more than the latter one, indicating that it's a classic imbalanced data set.

In order to classify the imbalanced data set about Leukemia, current classification methods are investigated. Based on the problems of sampling rare and small dataset, a mixed-sampling classification method is proposed, in which decision tree is used as base classifier. Experiments are performed to verify the correctness and feasibility of this method.

# 3 Data Imbalanced Problem

Since the data in real-world applications is often imbalanced, many machine leaning applications involving imbalanced data sets have been developed for solving the problem of imbalanced classification. There are two main types of strategies for solving the imbalanced classification problems: the first is to extract a balanced training set, and the second is related to the improved classification algorithm. The first method focused on the way of changing the imbalanced training data set into balanced training data set by reducing the scale of the majority class or enlarging the minority class. There are three types of balanced training set method: over-sampling of training set, under-sampling of training set and partition training set. The improved classification algorithm method tries to modify the classification algorithm to fit the imbalanced training data set. It includes ensemble classifier, Cost-sensitive learning and the features selection etc.

In some cases, the methods related to the improved classification algorithm are very effective, but have the disadvantage of depending on the specific algorithms [27]. As for data sets of different characteristics, the algorithm suitable for their classification is different, that is, every classification algorithm is generally a good fit for certain data sets, and its application in other data sets would be difficult [28]. However, sampling method is independent of specific algorithms. Strictly speaking, it is a data pre-processing method. Sampling can improve the balance of data set. And based on data set being sampled, different classification algorithm can be used for modeling in more flexible way. Therefore, in this paper, based on the investigation of sampling methods, a mixed-sampling method is proposed to solve the problem in classifying the imbalance data set.

Many researchers have developed the methods for under-sampling. Tomek used the distances of samples to create Tomek links, and deleted the Tomek links in majority classes to reduce the noise and the border sample[29]. Hart provided a Condensed Nearest Neighbor Rule (CNN) for under-sampling[30]. Kubat and Matwin provided an One Sided Selection (OSS) for under-sampling by integrating the Tomek links and CNN[31]. Other works, including Neighborhood Cleaning Rule (NCL)[32] , cluster-based under-sampling[33], inverse random under-sampling[34], are also done by researchers.

Partition is a typical solution for under-sampling. In this method, the samples of majority classes are divided into a series of data subsets without overlapping, whose scale is decided by the number of rare and small samples as well as pre-training sample distribution ratio. Then the subsets will be integrated with minority classes to form a series of balanced training data subsets, each of which can be trained as a base classifier. At last, the outputs of

those base classifiers are combined to form an ensemble classifier through meta-learning unit[39]. Polat, K. and S. Günes used C4.5 and one-against-all to solve multi-class classification problems[40].

Much work about over-sampling have been done too, including Synthetic Minority Over-Sampling Technique (SMOTE) [35], safe level SMOTE[36], noise replication in minority class[37]. Raskutti and Kowalczyk considered both over-sampling and under-sampling to evaluate the imbalance, and used only majority class training set or minority class training set to train the classifier only[38].

# 4 The approach

## 4.1 Problem Description

The sampling methods for imbalanced training data set including over-sampling and under-sampling have some limitations, such as random under sampling, random over sampling. Under-sampling may discard potentially useful data while over-sampling may increase the likelihood of over-fitting[41]. Despite these limitations, under-sampling and over-sampling in general are among the most popular sampling techniques and provide competitive results when compared with most complex methods[42].

In addition, partition of training data set is to get a series of balanced subsets. However, if there are small minority classes, the training data set may be parted into too small subsets, which might result in poor classification efficiency.

In this paper, mixed-sampling classification method is proposed to solve the problems for sampling imbalance data such as over-fitting in over-sampling, lost of potentially useful information in under-sampling and too small subsets resulting from partition of training data.

## 4.2 C4.5 Base Classifier - Decision Tree

In this paper, the C4.5 algorithm[43] proposed by Ross Quinlan is selected as the algorithm for base classifier.C4.5 becomes a very popular decision tree as based classifier to solve classification and regression task since it was developed in 1993[44]. It has additional features such as handling missing values, categorizing continuous attributes, pruning decision trees, rule derivation, and et al. C4.5 constructs a very big tree by considering all attribute values and finalizes the decision rule by pruning. It uses a heuristic approach for pruning based on the statistical significance of splits[45].

The testing attributes are selected according to information gain ratio in C4.5 algorithm which is described as follows.

The sample data set to be classified is set as $S = \{X_1, X_2, ..., X_n\}$ , where $X_i$ is a vector containing $m$ attribute items, $X_i = (A_1, A_2, ..., X_m)^T \ (1 \le i \le n)$.

Suppose the $A_m$ has $k$ different values, and the training data set $S$ could be divided into $k$ subsets, named as $C_1, C_2, \ldots, C_k$. The average information content of sample $S$ on the classification can be calculated by:

$$H(S) = -\sum_{p=1}^{k} P(C_p) \log_2 P(C_p)$$

$$\text{where } P(C_p) = \frac{|C_p|}{|S|} \quad (1 \le \mathrm{p} \le \mathrm{k})$$

The process of constructing decision trees is the one to make the uncertainty less and less after classification. Taking discrete attribute $A_i (1 \le i \le m-1)$ as an example, suppose it has $t$ different values named $a_q (1 \le q \le t)$, according to the values of $A_i$, not only sample set $S$ can be divided into $t$ subset, but also can $C_1, C_2, \ldots, C_k$ subsets be divided into $k \times t$ subsets named $C_{pq}$ ( $1 \le p \le k$ , $1 \le q \le t$ ), each of which represents the sample set which belongs to $p$ class under the condition of $A_i = a_q$. Then, information content of sample $S$ on the classification corresponding to attribute $A_i$ can be calculated by:

$$H(S / A_i) = -\sum_{q=1}^{t} P(C_q) \left[ -\sum_{p=1}^{k} (C_{pq}) \log_2 P(C_{pq}) \right]$$

$$\text{where } P(C_q) = \sum_{p=1}^{k} \frac{|C_{pq}|}{|S|}, \ P(C_{pq}) = \frac{|C_{pq}|}{|S|},$$

Then $Gain(S, A_i)$, the information gain resulting from classifying sample $S$ by using attribute $A_i$, represents the degree of uncertainty declination after $S$ is partitioned. The formula is as follows.

$$Gain(S, A_i) = H(S) - H(S / A_i)$$

The ratio of information gain is equal to the information gain divided by the split information content. And its formula is as follows.

$$GainRatio(S, A_i) = \frac{Gain(S, A_i)}{SplitGain(S, A_i)}$$

where $SplitGain(S, A_i) = -\sum_{g=1}^{t}\left(\frac{|S_g|}{|S|}\right)\log_2\left(\frac{|S_g|}{|S|}\right)$

Not only discrete attributes but also continuous attributes can be processed with C4.5 algorithm. The basic idea is to divide continuous value into a set of discrete values.

Before C4.5 is applied, it is necessary to identify two parameters of decision trees: optimal leaf size and importance of attributes, thus to improve the efficiency of C4.5. Processes of identifying them are *Finding the Optimal Leaf Size* and *Estimating Feature Importance*.

### 4.3 Ensemble Method

In the field of machine learning and pattern recognition, the ultimate goal is to achieve the best performance of classification. In traditional approach, different classification methods are tried to solve the target problem, and the classifier which has the best classification efficiency would be selected as the final resolution. However, it is found that the samples which are incorrectly classified by different classifiers are not entirely overlapped. That is, the sample classified wrongly by one classifier may be corrected by another, suggesting that different classifiers can provide complementary information to improve the classification efficiency. That is why researchers focus on integrating complementary information from different classifiers to improve the classification efficiency, which is origin of idea of ensemble classifiers. Ensemble classifiers can get better classification results by integrating different outputs of different classifiers[46, 47].

Krogh, A. and P. Sollich provided a general definition of ensemble learning in 1997:

A finite number of classifiers are applied to learn the same problem, and the output of ensemble classifiers is decided by outputs of all classifier [48].

Generally ensemble learning can be divided into learning phase and application phase. As shown in Figure 2, in learning phase, the *k* training data sets $TS_i\ (i = 1,2,...,k)$ are generated from original training data set. Each training data set $TS_i\ (i = 1,2,...,k)$ can generate a corresponding base classifier $h_i\ (i = 1,2,...,k)$. In application phase, base classifiers can be integrated into one ensemble learning system $h^* = F(h_1, h_2,..., h_k)$ in certain way, which will be applied in classifying testing sample *X*.

Class labels should be set for the outputs after ensemble learning. Suppose that the research object of ensemble learning system is a problem belonging to Category *c*, and *k* base classifiers are used. For each classifier $h_i\ (i = 1,2,...,k)$, the output is a *c* dimension vector, that is, $Output(h_i) = [d_{i,1}, d_{i,2},..., d_{i,c}]\ (i = 1,2,...k)$, where . $d_{ij}\ (i = 1,2,...,k;\ j = 1,2,...,c) \in [0,1]$. The value of $Output(h_i)$ represents the probability that the sample belongs to Category *j*. $h_i$ will distribute the sample into the category which has the maximum value in the output vector.

There are many methods for identifying the category of ensemble outputs, such as maximum method, maximum and minimum method and mean method. In this paper, mean method is selected to categorize the ensemble outputs. Suppose $\mu_j(x)\ (j = 1,2,...,c)$ is the probability that sample *x* belongs to Category *j*, sample *x* would be distributed into the category with maximum value of $\mu_j(x)$ which can be calculated by:

$$\mu_j(x) = \frac{1}{k}\sum_{i=1}^{k} d_{ij}\ (j = 1,2,...,c)$$

In summary, decision trees are used as base classifiers and mean method is selected for ensemble decision for categorization. In ensemble classification, it's necessary to estimate the number of decision trees, enabling ensemble classification systems to achieve better Price-Performance Ratio.

### 4.4 Mixed-Sampling

As for classification issues, the training data will significantly influence the classification accuracy. However, the data in real-world applications are often imbalanced class distribution. In this case, if all the data are used as training data, classifiers tend to bias against the minority class. Hence, it is important to select the suitable training data for classification in the imbalanced class distribution problem. We try to integrate over-sampling minority class and under-sampling majority class into mixed-sampling for selecting more suitable training data for classification.

The process of mixed-sampling proposed in this paper is shown in Figure 3. The steps are listed in details as follows.

Step 1: The imbalanced data set about Leukemia is divided into training data set and testing data set with the ratio of 2:1. The mixed-sampling is performed only on training data set, while the testing data set is used for evaluating the efficiency of classifiers.

Step 2: Training data set is divided into three categories according to the scale of classes, such as majority class, middle class and minority class. Different methods are applied in sampling these different categories.

Step 3: As for training data sets of majority classes, they are partitioned into subsets those have similar scale as middle class.

Step 4: As for training data sets of minority classes, they are processed to form subsets those have similar scale as middle class with over-sampling with replacement.

Step 5: All of training data sets and subsets with similar scale are combined with full permutation, thus to form a series of balanced training data subsets.

Step 6: Base classifiers with decision trees are used to learn with balanced training data subsets formed in the previous steps.

Step 7: These base classifiers are integrated with categorization decision method based on mean value, and an ensemble classification system is constructed with mixed-sampling.

Step 8: Imbalanced test data set is used to test and evaluate the ensemble classification system to verify its efficiency.

### 4.5 Evaluation Method

As the main evaluation criteria of classifiers, precision rate represents the percentage of the samples correctly classified in all samples. But to some extent, it may evaluate the distribution of classification results but ignoring the cost for wrong classification, especially in imbalanced data set, in which the cost of wrong classification in minority classes is much more significant than that in majority classes. For example, there are Class A and B in imbalanced training data set (IM). There are 99 data in A and only 1 in B. Suppose a sample wrongly classified belongs to Class A, the precision rate of Class A is 98.99% and that of Class B is 100%. However, if this sample belongs to Class B, the precision rate of Class A is 100%, and that of Class B is 0, indicating a much more significant cost in Class B.

To resolve the deficiency in precision rate, Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) attract much more concerns, for they can be used to evaluate the classifier for two-class issue. ROC and AUC not only consider the concurrency of errors, but also can distinguish the cost for these classification errors, which present efficiency performance in evaluating the classifier for two-class issue [49].

In order to draw ROC curve, confusion matrix should firstly be defined as shown in Table 4, in which P is positive class, N is negative class, T is true classification and F is false classification while TP is the number of samples classified correctly into positive class, FP is the number of samples classified wrongly into negative class, FN is the number of samples classified wrongly into negative class and TN is the number of samples classified correctly into negative class.

From confusion matrix in Table 4, the probability matrix of two classes is obtained as shown in Table 5.

Here,

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

$$FPR = \frac{FP}{FP + TN}$$

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

$$TNR = \frac{TN}{FP + TN}$$

Using FPR as X-axis and TPR as Y-axis, ROC curve is drawn in the coordinate system, on which a series of points (X,Y) is obtained by adjusting the threshold of decision strategy in classifier. the more convex ROC curve is, and closer it is to the upper left corner, the stronger the capacity for classification is. AUC is the area under ROC curve[50], that is, the AUC is the integral of ROC curve. The value of AUC belongs to [0, 1] while generally it belongs to [0.5, 1]. The value of AUC can be used to evaluate the classification capacity of classifier, the larger the value, the stronger the classification capacity.

Although ROC and AUC are only fit for evaluating two-class issue, it can be used to evaluate multi-class issue after it is processed[51]. How to use AUC to address multi-class issue still lacks a standard method. In this paper, multi-class issue will be transformed into many two-class issues, for each of which ROC curve is drawn and AUC is calculated respectively. And the mean value of all two-class issues is taken as the AUC of multi-class issue.

Suppose there are c categories in multi-class issue, one of which is named g（$1 \le g \le c$）and selected as positive class, and the rest as negative class. Then

the value of $AUC_g$ is calculated. And the AUC value in c categories is calculated by:

$$AUC = \frac{1}{c}\sum_{g=1}^{c} AUC_g$$

Thus, the value of AUC can be used to evaluate the classification capacity of the classifier. Similarly, the larger the AUC value, the stronger the classification capacity.

# 5 Experiments

## 5.1 Features of Leukemia's Literature

In PubMed, every journal article is indexed with about 10 to 15 subject headings or subheadings which represent the main content of the article. 2093 subject headings in the data set about common types of leukemia are statistically analyzed. Then, taking term frequency 10 as the threshold, 650 subject headings are preliminarily selected as the attribute dimension of literature vectors, thus to form a 2093 multiply 650 matrix where the row vector is observation vector and the column vector is attribute vector. Vector of common types of leukemia corresponding to PubMed literatures is taken as target classification and integrated with the observe matrix, and a 2093*651 matrix is formed where the first column vector is target vector which represents the classification identifier of each literature.

## 5.2 Parameters Values

Before training, there are three parameters in mixed-sampling to be confirmed, including the optimal leaf size of decision trees, the number of decision trees and the importance of sample attributes.

**a) Finding the Optimal Leaf Size**

To detect the optimal leaf size of decision trees, 50 base classifiers are used to construct the ensemble classifiers with different leaf sizes of 1, 5, 10, 20, 50 and 100, forming 6 different kinds of ensemble classifiers. Classification errors made in ensemble classifiers with 6 different leaf sizes are calculated and the result is drawn in Figure 4.

As we can see from Figure 4, the amount of classification errors is the minimum when the number of decision trees is larger than 12 and the leaf size is 1. Therefore, 1 is selected as the value of leaf size of decision trees. As 1 is the boundary value, classification errors of ensemble classifiers are calculated for comparison when leaf size is 1, 2, 3, 4 or 5. All curves about classification errors are shown in Figure 5. The amounts of classification errors in the five classifiers are close. However, when there is more and more decision trees involved, the classification errors would be less in ensemble classifiers when the leaf size is 1.

With all the experiments and comparisons mentioned above, 1 is identified as the value of leaf size of decision tree.

The value of leaf size parameter is set to 1 after all the experiments and comparisons above.

**b) The Number of Decision Tree**

The number of decision trees is closely related to the classification efficiency of ensemble systems. When the leaf size is 1, classification errors of ensemble classifiers consisting of different number of decision trees are calculated as drawn in Figure 6. The number of classification errors is the minimum when the number of decision trees is about 40. Therefore, 40 are set to be the parameter of the number of decision trees in ensemble classifiers.

**c) Estimating Feature Importance**

After preliminary processing, 650 subject headings are selected to be the features for classification. However, some of them don't contribute to classification and even derogate classification accruacy. Therefore, only the most important attributes should be selected to improve classification efficiency.

The data subsets are classified according to different attributes on condition that the leaf size is 1 and the number of decision trees is 40. The classification efficiency is shown in Figure 7, and the number of classification errors is shown in Table 6.

As shown in Table 6, 0.1338, the 19th figure is the minimum value of classification errors. And there are 190 attribute features (index multiply 10) selected to classify the data subsets to achieve least errors in classification. So, we select the first 190 features to perform the experiment.

The importance of 190 attribute features is evaluated and the result is shown in Figure 8.

## 5.3 Experiments

Research object in this paper involves Leukemia's data sets of 6 categories, which are AML, ALL, CML, CLL, AML+ALL and CML+CLL. These 6 categories are then divided into 6 sub categories, whose $TPR_g$, $FPR_g$ and $AUC_g$ (g=1,2,3,4,5,6) are calculated, where $AUC_g$ is calculated when the g class is positive class. Then, the total TPR, FPR and

AUC are calculated by averaging, which can be used to evaluate classification efficiency.

In order to eliminate causal factors and achieve unbiased results, 10 runs are performed at random for each experiment. The result and stability of classification is evaluated by averaging AUC in 10 calculations, and its stability is evaluated by Standard Deviation.

In the experiment, the formula of standard deviation is calculated by:

$$std = \left( \frac{1}{10} \sum_{g=1}^{10} (AUC_g - \overline{AUC})^2 \right)^{\frac{1}{2}}$$

$$\text{where} \quad \overline{AUC} = \frac{1}{10} \sum_{g=1}^{10} AUC_g$$

Before each experiment, the data sets should be divided into training data set and testing data set according to common method. This step will not be mentioned repeatedly in later experiments.

Eight classification methods are experimented and comparably analyzed, thus to verify the efficiency and stability of mixed-sampling method in classifying imbalance data set about Leukemia, which are listed as follows:

a) Single Decision Tree
b) Single Decision Tree with Over-sampling
c) Ensemble 40 Decision Trees
d) Ensemble 40 Decision Trees with Over-sampling
e) Partition
f) Partition and Over-sampling
g) Mixed-sampling
h) Ensemble 10 Decision Trees with Mixed-sampling

For the convenience of statement, these 8 methods are named by their sequence as Method a, b, c, d, e, f, g and h. The results and comparable analysis of the experiments for them are stated in detail as follows.

The classification result of single decision tree method is shown in Table 7. The total mean of AUC is only 0.678081, indicating a poor classification efficiency. And the AUC values of minority classes (CLL, AML+ALL and CML+CLL) are less than 0.61, indicating that their classification efficiency is close to that of random classification.

The basic reason for poor classification efficiency of Single Decision Tree is the imbalance of training data sets. To solve this problem, the strategy of sampling with replacement in the minority classes is applied to construct balanced training data sets, which are used to learn in single decision tree. The result is shown in Table 8.

As we can see from Table 8, the total mean value of AUC is only 0.693618, indicating the classification efficiency is still poor. However, it is better than that in Method a, and the AUC values of minority classes (CLL, AML+ALL and CML+CLL) have all elevated to above 0.61.

The basic reason for poor classification efficiency of Single Decision Tree is the imbalance of training data sets. Another way to solve this problem is to utilize ensemble classifier, performing different training on the imbalance data sets thus to achieve better efficiency. In Method c, base classifier with 40 decision trees and bagging method are used for ensemble learning. And its classification result is shown in Table 8.

As we can see from Table 8, the total mean value of AUC, 0.865027, is much larger than that in Method a and b (0.678081 and 0.693618), indicating that Method c has a much better classification efficiency compared with Method a and b. However, the AUC values of two minority classes （AML+ALL and CML+CLL） are less than 0.75 (0.712681 and 0.72821 respectively), showing that Method c do not have a good classification efficiency in minority classes. Furthermore, the standard deviation value of Class “CML+CLL”, 0.161885, is much higher than that of other classes, revealing that the classification efficiency of Method c in Class “CML+CLL” is not stable.

In order to improve classification efficiency and stability of Method c in minority classes, sampling with replacement to minority classes in training data set is used to balance the whole training data set before ensemble with bagging method. Then Method c is used to classify training data set, and the classification result is shown as Table 8.

As we can see from Table 8, the total mean value of AUC, 0.892858, is larger than that in Method c (0.865027), indicating that classification efficiency of Method d is better than that of Method c. Moreover, the AUC values of two minority classes (AML+ALL and CML+CLL) which are 0.799774 and 0.811507 respectively, are much larger than those of Method c (0.712681 and 0.72821). And both are larger than 0.75, and the standard deviation values of Class “CML+CLL” has decreased from 0.161885 to 0.811507, indicating that the classifiers in Method d are improved in terms of both classification efficiency and stability in Class ” CML+CLL”.

Therefore, compared with Method c, Method d achieves better classification efficiency in all classes and better stability in minority classes.

In order to improve classification efficiency of Method d, the imbalance in training data sets should

be further processed to achieve balance. The training data sets of majority classes are parted randomly into some subsets, which are integrated with minority classes in training data sets to form balanced training data sets. The number of samples for partitioning is set to 30, and the number of subsets in each class is listed in the Table 89.

The total number of data sets used for training is 780, which is calculated by full permutation (13*30*2*1*1*1). The base classifier with 780 decision trees are used in the ensemble learning system, and the classification result is shown as Table 8.

As we can see from Table 8, the total mean value of AUC (0.927361) is above 0.9 and much larger than that in Method c and d (0.865027 and 0.892858), indicating a good classification efficiency of Method e. Moreover, the AUC values of three minority classes (CLL, AML+ALL and CML+CLL), 0.984554, 0.963636 and 0.846821 respectively, are much larger than those in Method d which are 0.947323, 0.799774 and 0.811507 respectively.

As we can see from Table 8, relatively, the AUC value of majority class "ALL" is 0.872372, indicating that the classification efficiency of Method e is not good in this class. This may due to potential useful information lost caused by too many partitions which resolve the correlations in majority classes. Additionally, every training data set of three minority classes in Method e has to be trained for 780 times, which might result in over-fitting learning and decrease the classification efficiency of majority class "ALL".

In Method e, the number of partition subset is 30, which causes the number of every class in the training subset may be 30, 30, 18, 22 and 8. It is a generally balance data set but with certain imbalance. Over-sampling with replacement in minority classes is performed on the training subsets to form a subset with full balance, which will then be trained. The classification result of Method f is shown in Table 8.

As we can see from Table 8, the total mean value of AUC, 0.902226, is some less than that in Method e (0.927361), indicating worse classification efficiency of Method f. The reason is that over-fitting has already existed in Method e, and over-sampling with replacement in Method f intensifies over-fitting, thus influence the whole classification efficiency. In addition, it also causes the further decrease of classification efficiency in majority class "ALL", for the value of AUC has decreased from 0.872372 to 0.83697.

Therefore, Method f can't solve the problems about over-fitting and potential useful information lost existing in Method e.

Mixed-sampling method will be used to solve the problems about over-fitting and potential useful information in Method e. Firstly, the scale of training subsets in each class is set to 100. Then each training data set is partitioned into subsets and the number is shown in Table 810.

The total number of data subsets used for training is 36, which is calculated by full permutation (4*9*1*1*1*1). The base classifier with 36 decision trees is used in the ensemble learning system, and the classification result is shown as Table 8.

As we can see in Table 8, total mean value of AUC, 0.902687, is little less than that in Method e (0.927361), indicating the classification efficiency of Method g is not good as that of Method e. However, as for the ensemble classifiers involved, the number of base classifiers is only 36, which is only 4.62 percents of that in Method e which is 780.

In Method g, less base classifiers are used but achieve similar classification efficiency as Method e which involves large quantity of base classifiers. And the stability of these two methods is comparative. Method g has a much better performance than Method e.

In this method, AUC value of majority class "ALL" slightly changed from 0.872372 to 0.872374, indicating the classification efficiency of Method g in majority class "ALL" is not reduced.

In method g, only a small number of base classifiers are used. If the number of base classifiers in method g is increased, the classification efficiency would be much better. Here, the ensemble classification with 10 decision trees is performed on each training data set parted in the same way as in Method g. So the base classifiers in Method h reaches 360 (36*10), which is 46.2 percents of 780 base classifiers in Method e. The classification result of Method h is shown in Table 8.

As we can see from Table 8, total mean value of AUC, 0.933208, is little larger than that in Method e (0.927361), indicating a better classification efficiency of Method h than Method e but with only 46.2% of the number of base classifiers in Method e. Moreover, for majority class "ALL", the AUC value is 0.901947, suggesting a greatly improved efficiency. That is, the problems about over-fitting and potential useful information lost in Method e are to some extent solved by Method h.

## 5.5 Comparison and Result

The mean values and standard deviation values of AUC in 8 classification methods mentioned above

are shown as Table 811. As we can see from Table 811, there is not much difference between each value of "AUC Std.", indicating that 8 classification methods have the strong stability in efficiency, and the mean value of AUC can represent the stability of classification.

On the other hand, Method h has better classification efficiency than other methods, indicating that a small number of decision trees are integrated to effectively improve the classification efficiency with less cost of performance of the classifiers.

In summary, the mix-sampling classification method proposed in this paper can effectively improve efficiency of classifying the imbalanced data sets about Leukemia.

# 6 Conclusions and future works

In this paper, the data set about Leukemia is created by using the data of CGC, PubMed literatures and gene2pubmed, which is an imbalanced data set. The mixed-sampling classification method is proposed to classify this imbalanced data set, in which the decision tree with C4.5 algorithm is selected as based classifier, and the multi-classes issue is transferred to two-class issue. Indicators like ROC and AUC are used to evaluate the classification efficiency of this method.

Eight methods such as single decision tree, single decision tree with over-sampling are used to perform classification and testing experiments on data set about Leukemia. And indicators like ROC and AUC are used to evaluate its classification efficiency. Through experiments and comparable analysis, it is verified that the mixed-sampling method has better classification efficiency than sampling with replacement and partition of training data set, and to some extent solve the problems like information lost resulting from over-fitting and too many partitions.

The results of the experiments indicates that the mixed-sampling classification method is suitable for constructing classifiers of imbalanced data set, and better classification results would be achieved.

Although it is proved that the mixed-sampling method proposed is effective in classifying the imbalanced data set about Leukemia, its classification efficiency has not been verified in imbalanced data set of other fields, which would be our future research work.

In addition, before mixed-sampling, imbalanced data set about Leukemia such as independent Component Analysis (ICA) needs to be transformed to improve the differentiation of training subset and achieve better effect, which is also part of our future work

# Acknowledgement

Supported by National Science Foundation of China (61100106); the Natural Science Foundation of Fujian Province of China (No.2010J05137 and 2011J01360); the Fundamental Research Funds for the Central Universities (No.2010121038)

*References:*

[1] Bonnet, D. and Dick, J.E., *Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell*. Nature medicine, 1997. 3(7): p. 730-737.

[2] Champlin, R. and Gale, R.P., *Acute lymphoblastic leukemia: Recent advances in biology and therapy*. Blood, 1989. 73(8): p. 2051.

[3] 3. Sawyers, C.L., *Chronic myeloid leukemia*. N Engl J Med, 1999. 340(17): p. 1330-40.

[4] Gale, R.P. and Foon, K.A., *Chronic lymphocytic leukemia*. Annals of internal medicine, 1985. 103(1): p. 101.

[5] Segel, G.B. and Lichtman, M.A., *Familial (inherited) leukemia, lymphoma, and myeloma: an overview*. Blood Cells, Molecules, and Diseases, 2004. 32(1): p. 246-261.

[6] Zhang, W., et al., *B-cell activating factor and v-Myc myelocytomatosis viral oncogene homolog (c-Myc) influence progression of chronic lymphocytic leukemia*. Proceedings of the National Academy of Sciences, 2010. 107(44): p. 18956.

[7] Wojcik, I., et al., *Abnormalities of the P53, MDM2, BCL2 and BAX genes in acute leukemias*. Neoplasma, 2005. 52(4): p. 318-324.

[8] Pinto, A., et al., *c-fos oncogene expression in human hematopoietic malignancies is restricted to acute leukemias with monocytic phenotype and to subsets of B cell leukemias*. Blood, 1987. 70(5): p. 1450-1457.

[9] Lishner, M., et al., *The BCL-1, BCL-2, and BCL-3 oncogenes are involved in chronic lymphocytic leukemia:: Detection by fluorescence in situ hybridization*. Cancer genetics and cytogenetics, 1995. 85(2): p. 118-123.

[10] Hattori N, Fukuchi K, and T., N., *p53 Protein Expression in Chronic Myelomonocytic Leukemia-1 Correlates with Progression to Leukemia and a Poor Prognosis*. Acta Haematol, 2011. 125(4): p. 242-246.

[11] Chaussabel, D. and Sher, A., *Mining microarray expression data by literature profiling*. Genome Biology, 2002. 3(10): p. research0055.
[12] Diaz-Uriarte, R. and Alvarez de Andres, S., *Gene selection and classification of microarray data using random forest*. Bmc Bioinformatics, 2006. 7(3).
[13] Peng, Y.H., *A novel ensemble machine learning for robust microarray data classification*. Computers in Biology and Medicine, 2006. 36(6): p. 553-573.
[14] Liu, K.H. and Huang, D.S., *Cancer classification using Rotation Forest. Computers in Biology and Medicine*, 2008. 38(5): p. 601-610.
[15] Li, X., et al., *An ensemble method for gene discovery based on DNA microarray data*. Science in China Series C-Life Sciences, 2004. 47(5): p. 396-405.
[16] K.H. Liu and C.G. Xu, *A genetic programming-based approach to the classification of multiclass microarray datasets*, Bioinformatics, 25(3): p. 331 – 337, 2009.
[17] Futreal, P.A., et al., *A census of human cancer genes*. Nature Reviews Cancer, 2004. 4(3): p. 177-183.
[18] Santarius, T., et al., *A census of amplified and overexpressed human cancer genes*. Nature Reviews Cancer, 2010. 10(1): p. 59-64.
[19] Tomlins, S.A., et al., *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer*. Science, 2005. 310(5748): p. 644-648.
[20] Cui, Q., et al., *A map of human cancer signaling*. Molecular systems biology, 2007. 3(1): p. 1-13.
[21] stlund, G., Lindskog, M., and Sonnhammer, E.L.L., *Network-based identification of novel cancer genes*. Molecular & Cellular Proteomics, 2010. 9(4): p. 648-655.
[22] An, X.Y. and Wu, Q.Q., *Co-word analysis of the trends in stem cells field based on subject heading weighting*. Scientometrics, 2011. 88(1): p. 133-144.
[23] Thomas, P., et al., *GeneView–Gene-Centric Ranking of Biomedical Text*. Gene, 2010.
[24] Xu, H., et al., *Gene symbol disambiguation using knowledge-based profiles*. Bioinformatics, 2007. 23(8): p. 1015.
[25] Haferlach, T., et al., *Global approach to the diagnosis of leukemia using gene expression profiling*. Blood, 2005. 106(4): p. 1189.
[26] Pagano, L., et al., *The epidemiology of fungal infections in patients with hematologic malignancies: the SEIFEM-2004 study*. Haematologica, 2006. 91(8): p. 1068.
[27] Estabrooks, A., Jo, T., and Japkowicz, N., *A multiple resampling method for learning from imbalanced data sets*. Computational Intelligence, 2004. 20(1): p. 18-36.
[28] Weiss, S.M. and Kapouleas, I., *An empirical comparison of pattern recognition, neural nets and machine learning classification methods*. Readings in machine learning, 1990.
[29] Tomek, I., *Two modifications of CNN*. IEEE Trans. Syst. Man Cybern., 1976. 6: p. 769-772.
[30] Hart, P., *The condensed nearest neighbor rule (corresp.)*. Information Theory, IEEE Transactions on, 1968. 14(3): p. 515-516.
[31] Kubat, M. and Matwin, S. *Addressing the curse of imbalanced training sets: one-sided selection*. 1997: Citeseer.
[32] Laurikkala, J., *Improving identification of difficult small classes by balancing class distribution*. Artificial Intelligence in Medicine, 2001: p. 63-66.
[33] Yen, S.J. and Lee, Y.S., *Cluster-based under-sampling approaches for imbalanced data distributions*. Expert Systems with Applications, 2009. 36(3): p. 5718-5727.
[34] Tahir, M., et al., *A multiple expert approach to the class imbalance problem using inverse random under sampling*. Multiple Classifier Systems, 2009: p. 82-91.
[35] Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique*. Journal of Artificial Intelligence Research, 2002. 16(1): p. 321-357.
[36] Bunkhumpornpat, C., Sinapiromsaran, K., and Lursinsap, C., *Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem*. Advances in Knowledge Discovery and Data Mining, 2009: p. 475-482.
[37] Lee, S.S., *Noisy replication in skewed binary classification*. Computational statistics & data analysis, 2000. 34(2): p. 165-191.
[38] Raskutti, B. and Kowalczyk, A., *Extreme re-balancing for SVMs: a case study*. ACM SIGKDD Explorations Newsletter, 2004. 6(1): p. 60-69.
[39] Masud, M., et al., *A multi-partition multi-chunk ensemble technique to classify concept-drifting data streams*. Advances in Knowledge Discovery and Data Mining, 2009: p. 363-375.
[40] Polat, K. and Günes, S., *A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems*. Expert

Systems with Applications, 2009. 36(2): p. 1587-1592.
[41] Batista, G.E., Prati, R.C., and Monard, M.C., *A study of the behavior of several methods for balancing machine learning training data*. ACM SIGKDD Explorations Newsletter, 2004. 6(1): p. 20-29.
[42] Liu, X.Y., Wu, J., and Zhou, Z.H., *Exploratory undersampling for class-imbalance learning*. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 2009. 39(2): p. 539-550.
[43] Quinlan, J.R., *C4. 5: programs for machine learning*. 1993: Morgan Kaufmann.
[44] Ripley, B., *Classification and regression trees*. R package version, 2005: p. 1.0-19.
[45] Utgoff, P.E., Berkman, N.C., and Clouse, J.A., *Decision tree induction based on efficient tree restructuring*. Machine Learning, 1997. 29(1): p. 5-44.
[46] Polikar, R., *Ensemble based systems in decision making.* Circuits and Systems Magazine, IEEE, 2006. 6(3): p. 21-45.
[47] Dietterich, T., *Ensemble methods in machine learning*. Multiple Classifier Systems, 2000: p. 1-15.
[48] Krogh, A. and Sollich, P., *Statistical mechanics of ensemble learning.* Physical Review E, 1997. 55(1): p. 811.
[49] Hand, D.J. and Till, R.J., *A simple generalisation of the area under the ROC curve for multiple class classification problems.* Machine Learning, 2001. 45(2): p. 171-186.
[50] Bradley, A.P., *The use of the area under the ROC curve in the evaluation of machine learning algorithms.* Pattern Recognition, 1997. 30(7): p. 1145-1159.
[51] Ferri, C., Hernández-Orallo, J., and Salido, M.A., *Volume under the ROC surface for multi-class problems*. Machine Learning: ECML 2003, 2003: p. 108-120.
[52] K.H. Liu, B. Li, Q.Q. Wu, et, al., *Microarray data classification based on ensemble independent component selection*, Computers in Biology and Medicine. 39(11): p. 953 – 960, 2009.
[53] K.H. Liu, B. Li, J. Zhang and J.X. Du, *Ensemble component selection for improving ICA based microarray data prediction models*, Pattern Recognition, 42(7): p. 1274–1283, 2009.

Table 1 Samples of CGC

| Symbol | Name | GeneID | Tumour Types (Somatic Mutations) | Tumour Types (Germline Mutations) |
|---|---|---|---|---|
| ABL2 | v-abl Abelson murine leukemia viral oncogene homolog 2 | 27 | AML | |
| AF15Q14 | AF15q14 protein | 57082 | AML | |
| AF1Q | ALL1-fused gene from chromosome 1q | 10962 | ALL | |
| FANCC | Fanconi anemia, complementation group C | 2176 | | AML, leukemia |
| AF3p21 | SH3 protein interacting with Nck, 90 kDa (ALL1 fused gene from 3p21) | 51517 | ALL | |
| BCL3 | B-cell CLL/lymphoma 3 | 602 | CLL | |
| BCL5 | B-cell CLL/lymphoma 5 | 603 | CLL | |
| MSI2 | musashi homolog 2 (Drosophila) | 124540 | CML | |
| HOXA11 | homeo box A11 | 3207 | CML | |
| FLT3 | fms-related tyrosine kinase 3 | 2322 | AML, ALL | |
| BCR | breakpoint cluster region | 613 | CML, ALL, AML | |
| … | … | … | … | … |

Table 2 Format of Gene2pubmed Text File

```
#Format: tax_id GeneID PubMed_ID
……
9       12465029873079
9       12465059873079
……
139     13430459593780
139     134304510678977
……
```

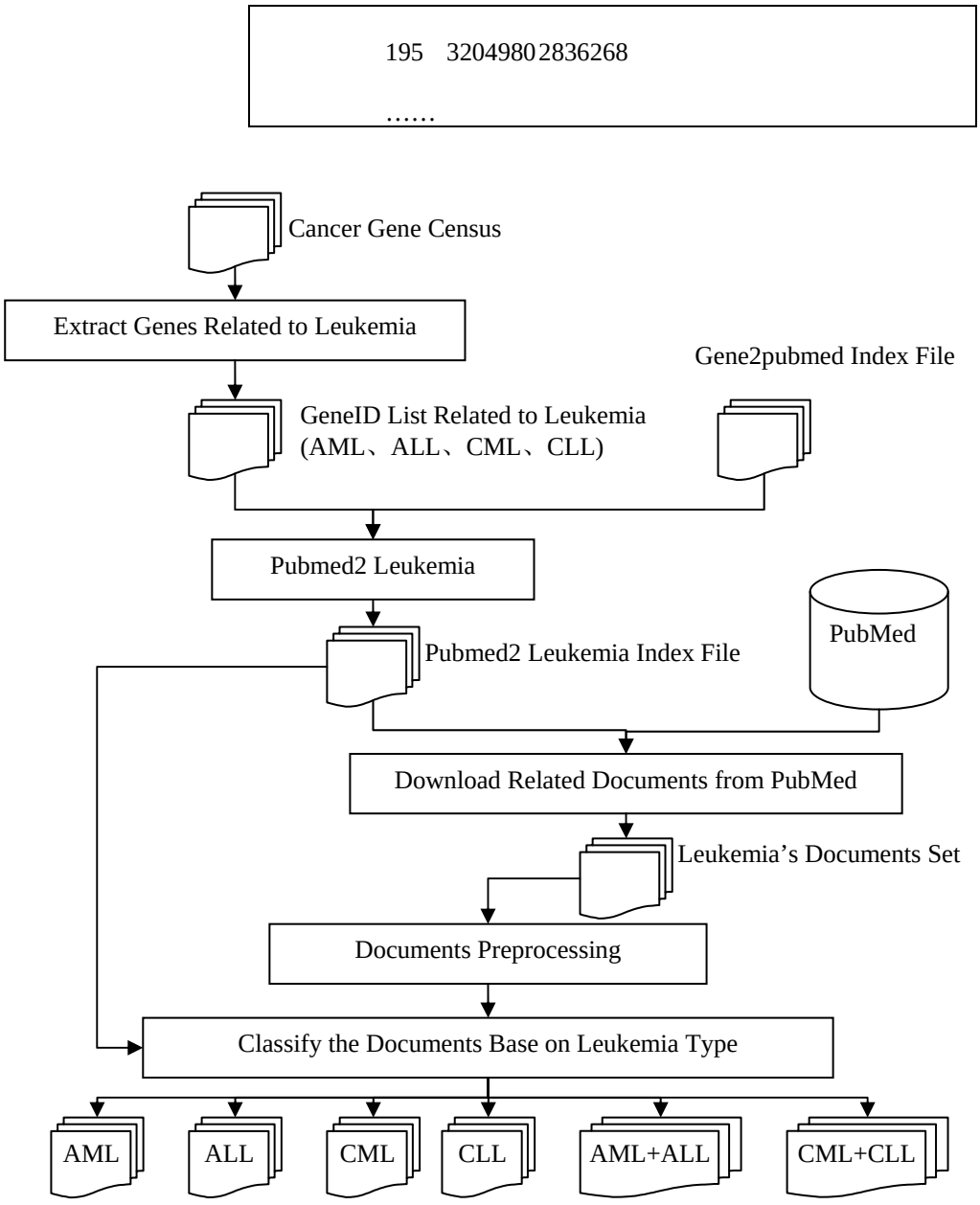


Figure 1 The Processes of Leukemia Documents Classification

Table 3 The Literature Distribution of Leukemia's Type

| Leukemia's Type | AML | ALL | CML | CLL | AML+ALL | CML+CLL |
|---|---|---|---|---|---|---|
| Num. of Literature | 580 | 1369 | 70 | 28 | 34 | 12 |

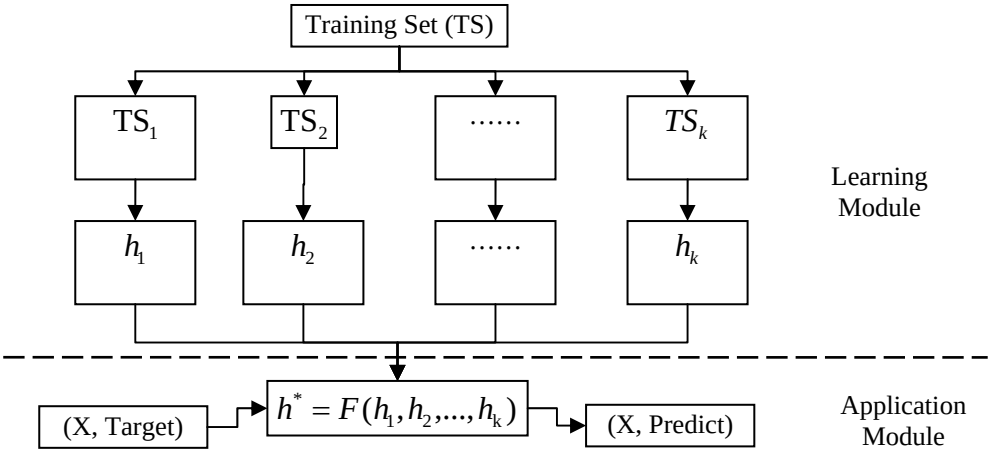


Figure 2 Ensemble Learning Diagram

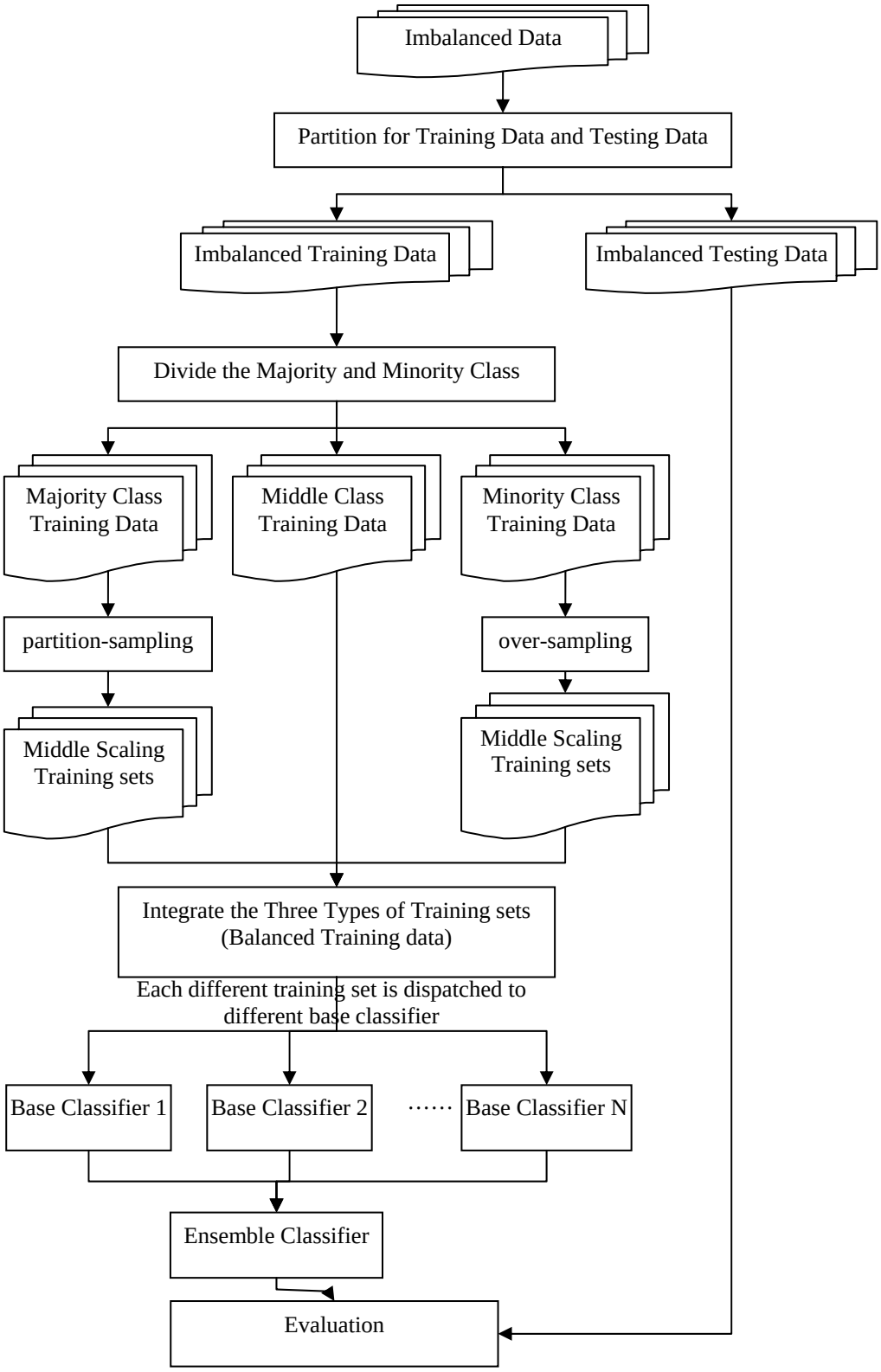


Figure 3 The Process of Mixed-sampling for Imbalanced Distribution

Table 4 Confusion Matrix of Two Classes

| Classifier classification | Expert classification | |
|---|---|---|
| | P | N |
| T | TP | FP |
| F | FN | TN |

Table 5 Probability Matrix of Two Classes

| | P | N |
|---|---|---|
| T | TPR | FPR |
| F | FNR | TNR |

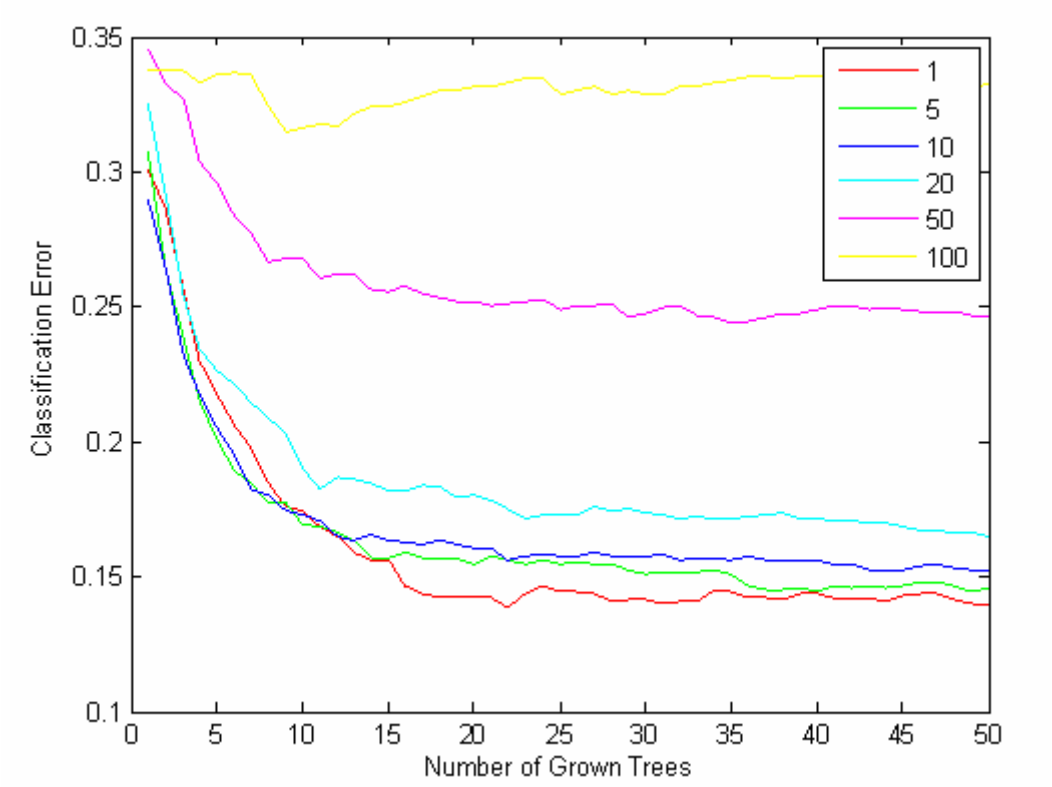


Figure 4 Classification Errors corresponding to different Leaf Sizes (1, 5, 10, 20, 50 and 100)

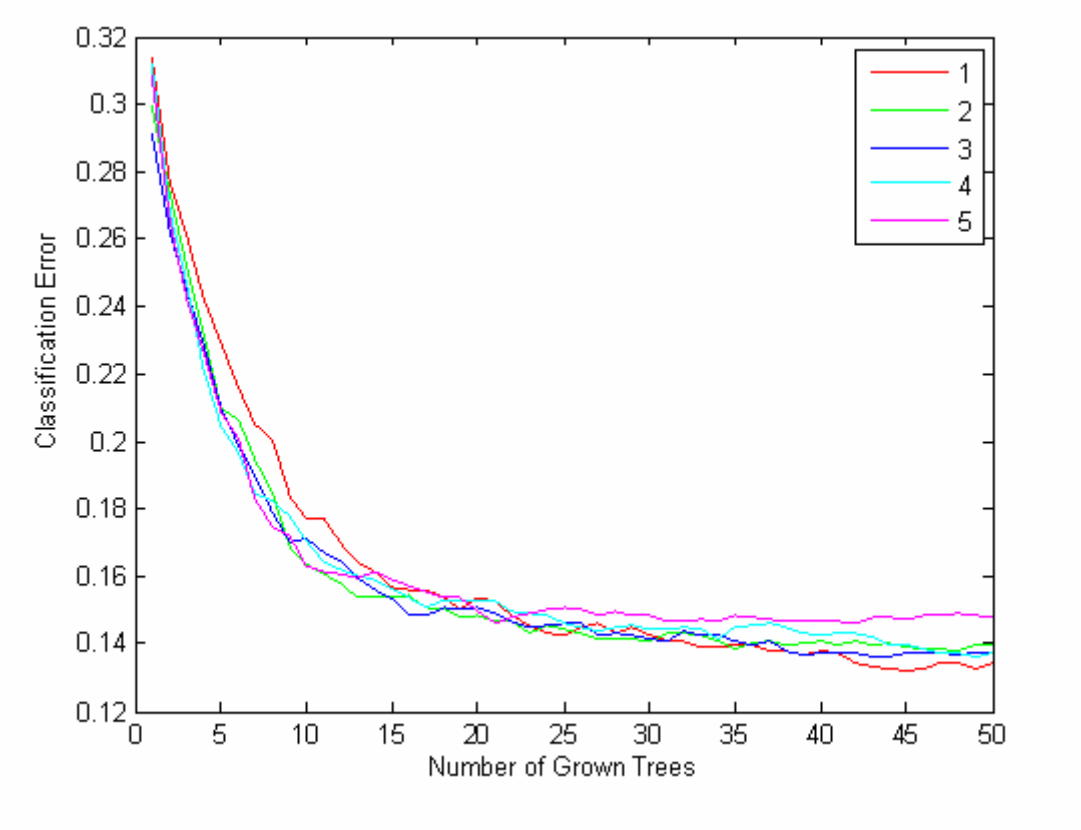


Figure 5 Classification Error to Leaf Size (1, 2, 3, 4 and 5)

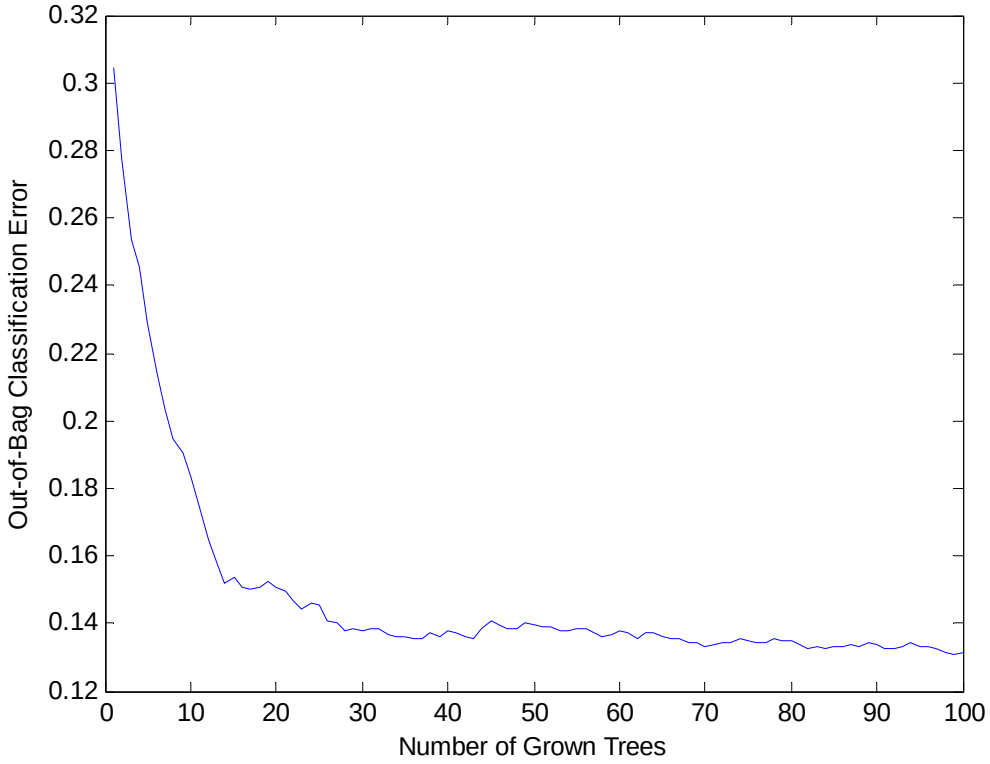


Figure 6 Classification Error of Grown Trees

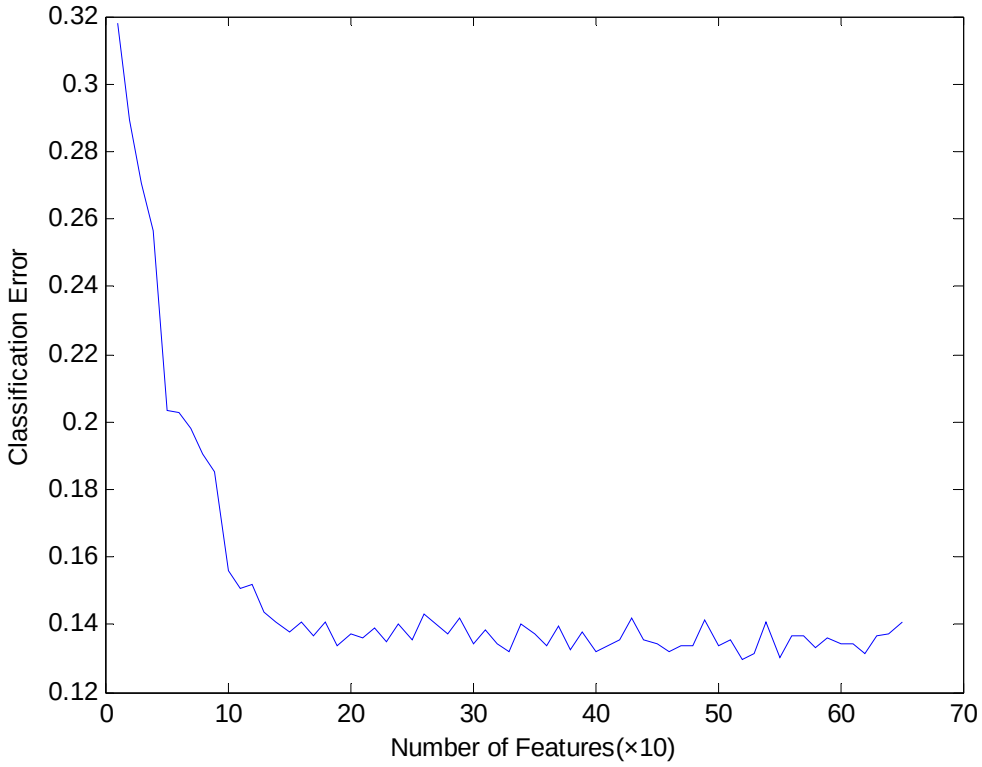


Figure 7 Classification Errors for Different Features

Table 6 Values of Classification Error (Indexed by Row)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.3182 | 0.2895 | 0.2704 | 0.2566 | 0.2035 | 0.2026 | 0.1983 | 0.1906 | 0.1854 | 0.1558 |
| 0.1505 | 0.1519 | 0.1438 | 0.1405 | 0.1376 | 0.1409 | 0.1366 | 0.1405 | **0.1338** | 0.1371 |
| 0.1362 | 0.1390 | 0.1347 | 0.1400 | 0.1352 | 0.1429 | 0.1400 | 0.1371 | 0.1419 | 0.1343 |
| 0.1386 | 0.1343 | 0.1319 | 0.1400 | 0.1371 | 0.1338 | 0.1395 | 0.1328 | 0.1376 | 0.1319 |
| 0.1338 | 0.1357 | 0.1419 | 0.1357 | 0.1343 | 0.1319 | 0.1338 | 0.1338 | 0.1414 | 0.1338 |
| 0.1352 | 0.1295 | 0.1314 | 0.1405 | 0.1304 | 0.1366 | 0.1366 | 0.1333 | 0.1362 | 0.1343 |
| 0.1343 | 0.1314 | 0.1366 | 0.1371 | 0.1409 | | | | | |

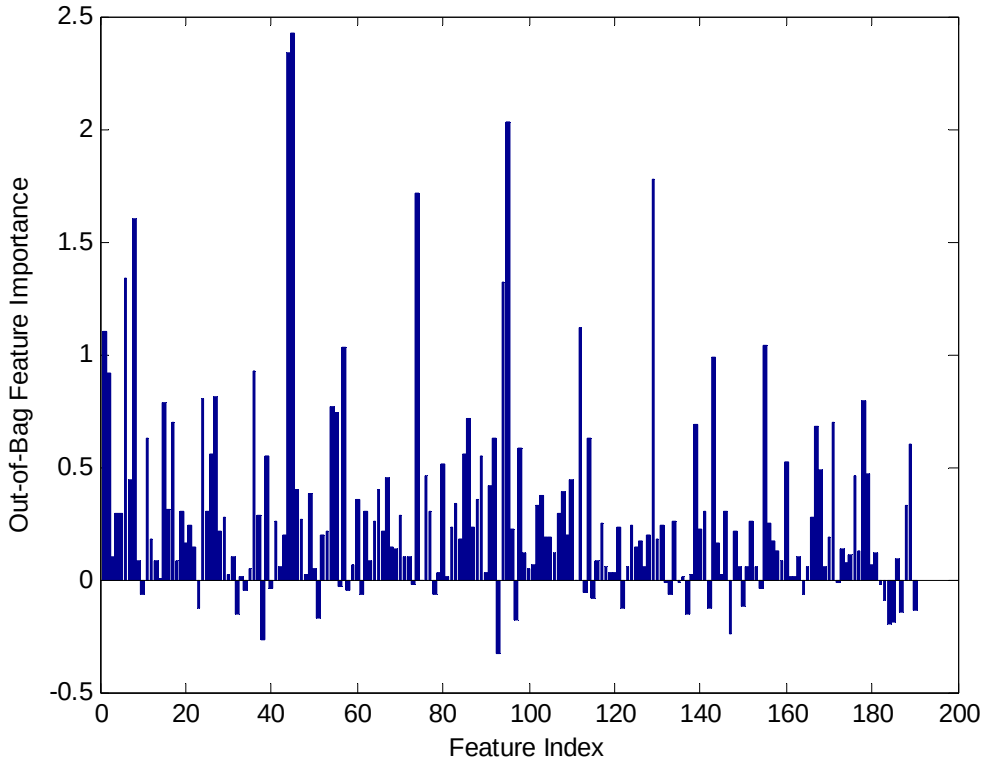


Figure 8 Importance of different attribute features

Table 7 Classification Result of Single Decision Tree

| Loops | AML | ALL | CML | CLL | AML+ALL | CML+CLL | Mean of One Loop |
|---|---|---|---|---|---|---|---|
| 1 | 0.812524 | 0.768512 | 0.671943 | 0.543911 | 0.575050 | 0.490607 | 0.643758 |
| 2 | 0.818967 | 0.766091 | 0.835196 | 0.594533 | 0.487591 | 0.616510 | 0.686481 |
| 3 | 0.757574 | 0.736225 | 0.759158 | 0.656720 | 0.701924 | 0.617775 | 0.704896 |
| 4 | 0.842144 | 0.795952 | 0.792816 | 0.596879 | 0.628534 | 0.489884 | 0.691035 |
| 5 | 0.813770 | 0.804258 | 0.752568 | 0.548278 | 0.577571 | 0.492775 | 0.664870 |
| 6 | 0.802377 | 0.784727 | 0.925673 | 0.593482 | 0.491971 | 0.489162 | 0.681232 |
| 7 | 0.783594 | 0.743937 | 0.796046 | 0.655507 | 0.561380 | 0.495665 | 0.672688 |
| 8 | 0.854345 | 0.803554 | 0.796725 | 0.602863 | 0.527074 | 0.494220 | 0.679797 |
| 9 | 0.760247 | 0.727664 | 0.752277 | 0.706130 | 0.515262 | 0.616329 | 0.679651 |
| 10 | 0.814641 | 0.780510 | 0.751728 | 0.526363 | 0.564167 | 0.621026 | 0.676406 |
| Mean | 0.806018 | 0.771143 | 0.783413 | 0.602466 | 0.563052 | 0.542395 | 0.678081 |
| std | 0.029905 | 0.026300 | 0.062677 | 0.053698 | 0.061853 | 0.061696 | 0.015353 |

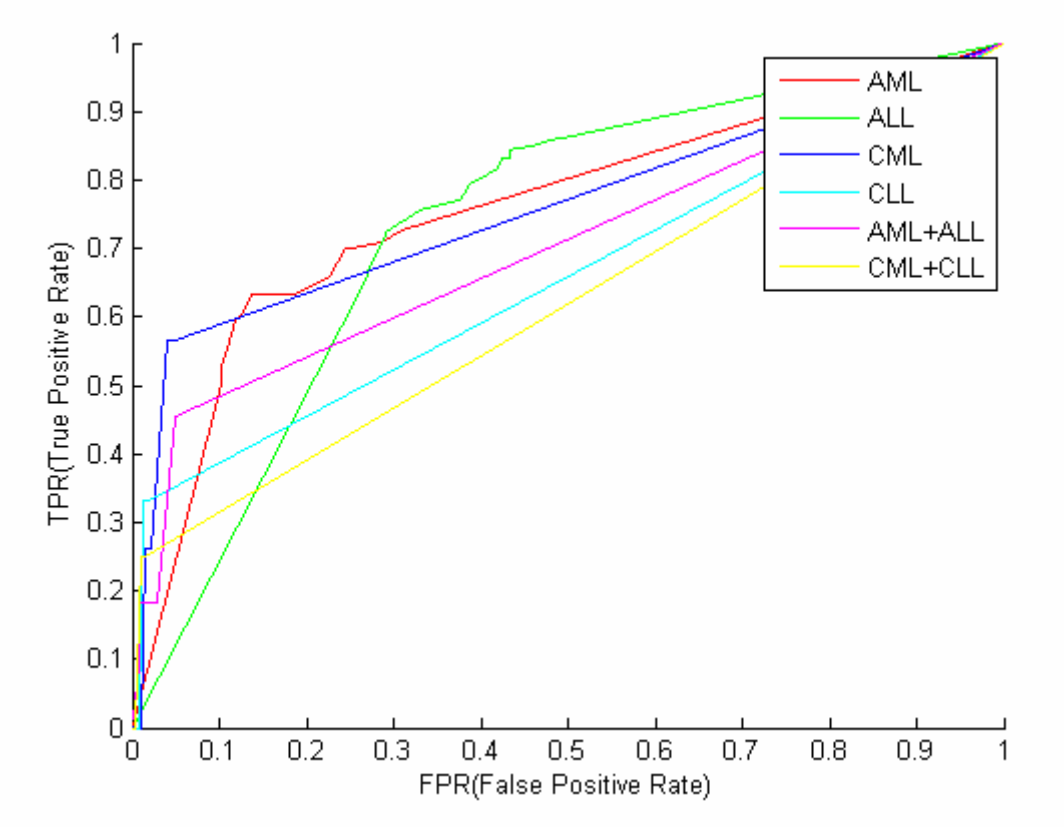


Figure 9 ROC curves of Single Decision Tree for different classes (Loops=3)

Table 8 Classification Result of Single Decision Tree with different methods

| **Method** | | **AML** | **ALL** | **CML** | **CLL** | **AML+ALL** | **CML+CLL** | **Mean of One Loop** |
|---|---|---|---|---|---|---|---|---|
| a | Mean | 0.806018 | 0.771143 | 0.783413 | 0.602466 | 0.563052 | 0.542395 | 0.678081 |
| | std | 0.029905 | 0.026300 | 0.062677 | 0.053698 | 0.061853 | 0.061696 | 0.015353 |
| b | Mean | 0.739553 | 0.704191 | 0.837651 | 0.629330 | 0.614585 | 0.636398 | 0.693618 |
| | std | 0.024971 | 0.024090 | 0.048954 | 0.050523 | 0.059757 | 0.109352 | 0.027014 |
| c | Mean | 0.938353 | 0.911452 | 0.974149 | 0.925311 | 0.712681 | 0.728215 | 0.865027 |
| | std | 0.007470 | 0.008755 | 0.013943 | 0.039327 | 0.063983 | 0.161885 | 0.031249 |
| d | Mean | 0.928667 | 0.895653 | 0.974226 | 0.947323 | 0.799774 | 0.811507 | 0.892858 |
| | std | 0.010068 | 0.011941 | 0.017690 | 0.017792 | 0.051476 | 0.097677 | 0.019724 |
| e | Mean | 0.929170 | 0.872372 | 0.967614 | 0.984554 | 0.963636 | 0.846821 | 0.927361 |
| | std | 0.007903 | 0.015414 | 0.016721 | 0.015884 | 0.024904 | 0.091759 | 0.018579 |
| f | Mean | 0.913245 | 0.836971 | 0.969636 | 0.980196 | 0.951387 | 0.761922 | 0.902226 |
| | std | 0.005300 | 0.016737 | 0.015611 | 0.015234 | 0.030577 | 0.134912 | 0.025256 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| g | Mean | 0.923221 | 0.872374 | 0.979017 | 0.985484 | 0.953928 | 0.702095 | 0.902687 |
| | std | 0.008239 | 0.012007 | 0.012465 | 0.013102 | 0.043893 | 0.070091 | 0.018027 |
| h | Mean | 0.942539 | 0.901947 | 0.984321 | 0.974252 | 0.942177 | 0.854010 | 0.933208 |
| | std | 0.007339 | 0.006732 | 0.014286 | 0.011586 | 0.033387 | 0.124018 | 0.022036 |

Table 9 Number of Partition Subset (Scale=30)

| Class Identifier | Data Subset Num. |
|---|---|
| AML | 13 |
| ALL | 30 |
| CML | 2 |
| CLL | 1 |
| AML+ALL | 1 |
| CML+CLL | 1 |

Table 10 Number of Partition Subset (Scale=100)

| Class Identifier | Data Subsets Num. |
|---|---|
| AML | 4 |
| ALL | 9 |
| CML | 1 |
| CLL | 1 |
| AML+ALL | 1 |
| CML+CLL | 1 |

Table 11 Comparisons among 8 Classification Methods

| **Method Order** | **Classification Method** | **AUC Means** | **AUC Std.** |
|---|---|---|---|
| a | Single Decision Tree | 0.678081 | 0.015353 |
| b | Single Decision Tree with Over-sampling | 0.693618 | 0.027014 |
| c | Ensemble 40 Decision Trees | 0.865027 | 0.031249 |
| d | Ensemble 40 Decision Trees with Over-sampling | 0.892858 | 0.019724 |
| e | Partition | 0.927361 | 0.018579 |

| | | | |
|---|---|---|---|
| f | Partition and Over-sampling | 0.902226 | 0.025256 |
| g | Mixed-sampling | 0.902687 | 0.018027 |
| h | Ensemble 10 Decision Trees with Mixed-sampling | 0.933208 | 0.022036 |