A Framework for Medical Text Mining using a Novel Categorical Clustering Algorithm

Anirban Chakrabarty
Department of MCA
Future Institute of Engineering and Management
Kolkata, India

ABSTRACT

The fast growth of medical records provides new opening for meaningful information retrieval in clinical diagnosis and treatment. Although nursing and pathology records provide a complete account of patient's information they are not fully utilized while taking major decisions of surgery or chemo therapy on patients. This research proposes a Minimum spanning tree algorithm to develop k-clusters of training data related to different liver diseases which are validated using Silhouette coefficient. A text classification algorithm is developed using cluster centers as training samples which uses a similarity measure to classify the categorical data. Simulation results show that the algorithm proposed can lower the calculation complexity and improve the accuracy of established text classification algorithms like k-NN. This research can serve as a medical diagnosis tool for classifying patient records and reveal important vocabularies that characterize nursing and pathology records.

Keywords: Categorical clustering; spanning tree; weight factor; silhouette coefficient; liver disease

1. INTRODUCTION

Text categorization is the task of assigning text document to one or more predefined categories according to its content and the labeled training samples. Automatic classification schemes can greatly facilitate the process of categorization [1]. However, given the fast growth in online document data, this would become more difficult with time. A relatively moderate size of document sets could easily have a vocabulary of tens of thousands of distinct words. Many existing algorithms simply would not work with these many number of attributes. Feature selection methods based on document frequency, mutual information, or information gain could be used to reduce the number of words[2]. Normally, the number of words after feature selection could be still in thousands.

There are several classification schemes that can be potentially used for text categorization however, many of these existing schemes do not work well in the text categorization. For example, widely used classification decision tree induction algorithm like C4.5 and RIPPER do not work well when the number of distinguishing features is large. Even though Naive-Bayes classification techniques, are commonly used in text categorization, however the independence assumption severely limits its applicability [3].

k-nearest neighbor (*k*-NN) classification is an instance-based learning algorithm that has shown to be very effective for a variety of problem domains including text classification. The key element of this scheme is the availability of a similarity measure that is capable of identifying neighbors of a particular document. A major drawback of the similarity measure used in *k*-NN is that it uses all features in computing distances[4]. Moreover the "nearest" measure used in k-NN is that it stores all the examples of the training set, creating high storage requirements and complexity in computing distance increases steeply as the dimension of data grows and also when large k values are considered for classification[5].

In this paper, a framework is proposed for automatically detecting a disease and classifying it from documents used at medical treatment sites. In the future, using a text data mining approach and laterally processing medical documents will support disease classification by retrieving the examples of similar syndromes. The work first develops k-clusters using a Minimum spanning tree algorithm (MSTCLUST) where each cluster contains patient dataset for a particular disease. The clustering algorithm sorts the edges of minimum spanning tree in descending order of weights and removes those edges from the tree that satisfy a predefined inconsistency measure. After building and validating the clusters, the work proposes a similarity weighted text classification algorithm (SWTCA) to classify a new patient document to an appropriate cluster.

The rest of the paper is organized as follows- Related work, Methodology, Proposed algorithms for cluster formation and algorithm for categorizing patients based on clusters weight. Finally experimental results and evaluation parameters are developed to determine the accuracy of the proposed framework.

2. RELATED WORK

Research work in text mining has generally focused either on text classification or named entity recognition e.g., finding keywords to classify the records or finding the terms of web documents, disease or syndrome. Several studies have been developed that are focused on biomedical literature mining. Reference[6] developed Medical Concept Mapper, a tool which maps synonyms and semantically related concepts. The system integrates NLP tools and ontologies and is designed to suggest cancer-related medical terminology based on user's query. Reference[7] identifies diseases and detects patients with a fever from free-text medical records. Work has been done to extract and mine a variety of information with breast complaints from free-text clinical records. This system proposes three approaches including ontology-based, graph-based, and NLP-based to perform text classification [8]. Several researchers have investigated the information extraction and knowledge discovery of biomedical literature by using ontologies. A system called BioPatentMiner, that facilitates information retrieval from

biomedical patents. This system first identifies biological terms and relations in the patents and then integrates information from these patents with biomedical ontologies and creates a biomedical semantic web [9].

Further research has been done using a bag-of-words approach to process the text of physical examination sections of in-patient and out-patient clinical notes in order to identify whether the findings of structural, neurological, and vascular components of a foot examination revealed normal or abnormal findings or if they were not assessed. A support vector machine classifier obtained accuracy of 88% for the vascular component [10].

Research has investigated the use of Minimum spanning tree for generating optimal clusters based image segmentation which is a fast and efficient method of generating a set of segments from an image. The algorithm uses a new cluster validation criterion based on the geometric property of data partition of the data set in order to find the proper number of segments. The algorithm works in two phases. The first phase of the algorithm creates optimal number of clusters/segments, where as the second phase of the algorithm further segments the optimal number of clusters/segments and detect local region outliers [11]. The process of text classification is assigning category c_i to document d_j for the predefined category collection{c1,c2,…,cm} and the given document collection{d1,d2,…,dn} and creating the mapping from collection D to collection C [12].

3. METHODOLOGY

3.1 Work Flow

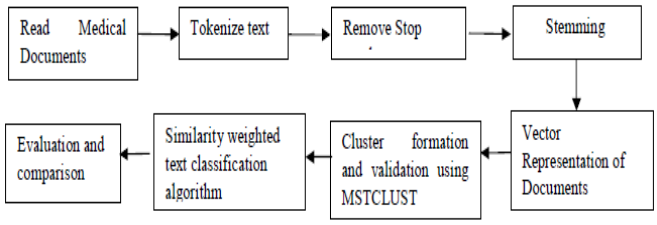

Figure-1: **Steps of Text categorization**

3.2 Descri

The purpose of this research is to develop an intelligent strategy that can assist in clinical decision-making process beyond traditional methods, based on an efficient utilization of the narratives in medical records for liver disease patients. The pre-processing for text classification includes many key technologies, such as sentence segmentation, remove stop words, stemming, feature extraction and weight calculation. This paper uses a vector space model to represent the text as the core idea is to make the document become a numeral vector in multi dimensional space.

The work develops k-clusters using a Minimum spanning tree algorithm (MSTCLUST) where each cluster contains patient dataset for a particular disease for eg: cluster-1 for liver cirrhosis. The MST clustering algorithm sorts the edges of minimum spanning tree in descending order of weights and remove those edges from the tree that are above the average weight. After building the clusters the work proposes a similarity weighted text classification algorithm(SWTCA) to classify a new patient document to an appropriate cluster. Feature weights are used in the similarity measure computation such that important features contribute more in the similarity measure.

3.3 Steps of the Similarity Weighted Text Classification Algorithm(SWTCA)

Input: Documents in category C_i ,{ di_1 ,di_2 , …,di_n }, the clusters S_i $i=1,..j$ when using SWTCA algorithm and the test sample document X.

Output: Document X's category is C_i .

Step1: Document X and all training samples are pre-processed-stop word removal, stemming done and the corresponding feature vector saved in excel file.

Step-2: Normalize the columns of Feature vector F dividing by highest word frequency to get the term frequency (TF) using formula –(1) mentioned as below.

$$TF_{ij} = Fij / \max(Fij). \quad (1)$$

Step-3: Among N documents, term j occurs in dj documents. Inverse document frequency measures uniqueness of term j. dj indicates the document frequency.

$$IDF_j = \log_2(N/dj) + 1 , dj > 0 \quad (2)$$

Step-4: Calculate weight of terms to determine their different significance

$$Wij = TF_{ij} \times IDF_j \quad (3)$$

Step-5: Select W_{ij} weights above a certain threshold Φ, and also remove zero values. This step removes terms which have less/no significance.

Step-6: Development of clusters is initiated by forming a patient to patient connected graph G.

Step-7: Using a novel minimum spanning tree cluster algorithm (MSTCLUST) generate the desired k clusters(for k disease sets) and then record the no. of samples in each cluster**.** Calculate the centers weight value of each category (or cluster) using the formula mentioned below.

$$y(Wij,Cj) = (\sum^{di}{}_{w=1} Wij)/ N_{Ci} \quad (4)$$

here Nci indicates the number of samples in category Ci .

Step-8: Use cluster centers as new training sets to classify the test document X. We calculate the probability of X belong to category Cj which has largest P(X,Cj).

$$P(X, C_j) = \sum_{W_{ij}\, c\, KNN} SIM(X, W_{ij}).\, y(W_{ij}, C_j) \quad (5)$$

Here the similarity measure $SIM(X,W_{ij})$ determines the similarity of test document X and training document W_{ij} and is based on Cosine Similarity Measure[13] calculated as

$$CSM(\bar{x}, \bar{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\sqrt{\sum_{i=1}^{n}(y_i)^2}} \quad (6)$$

Step-9: Judge document X to be the category which has largest P(X,Cj).

3.4 Minimum Spanning Tree Clustering Algorithm (MSTCLUST)

A spanning tree is an acyclic sub graph of a graph G, which contains all vertices from G and is also a tree. The minimum spanning tree(MST) of a weighted graph is the minimum weight spanning tree of that graph[14].

3.4.1Construction of Spanning Tree from Connected Graph:

Step1: Let E be an edge joining two patient documents in the connected graph G then the weight of an edge (W_e) is between each pair of document is calculated from Jaccard Index as shown below [15].

$$\rho(A,B) = (|A| \cap |B|) / (|A| \cup |B|) \quad (7)$$

here A and B are two patient documents.

Then for each pair find the distance measure between A and B .
$$dist(A,B) = 1 - \rho(A,B) \quad (8)$$

This measure can be used to calculate distance between categorical attributes [15].Thus the weight of an edge W_e between patient documents A and B is given by dist(A,B).

Step 2: Construct a connected weighted graph where the vertices are the patients and draw the edges between each pair of vertices where weight of each edge is the distance measure between the two corresponding patients represented as the vertices.

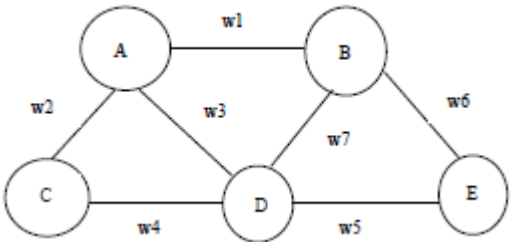

Figure2: **Shows a connected graph formed with patient documents**

Step 3: Identify the number of cycles in the graph.

For each cycle, remove the highest weighted edge involved in the cycle from the graph. If there exists more than one edge with the same highest value, choose any one of them arbitrarily and remove it from the graph.

At the end of this step a spanning tree is formed as shown in Figure-3.

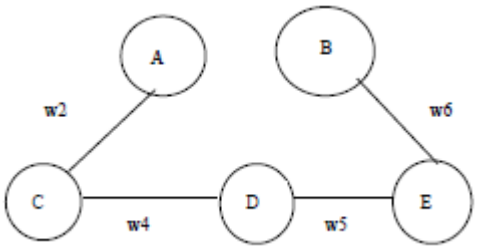

Figure-3: **A spanning tree formed by removing the cycles**.

3.4.2 Formation of Minimum Spanning Tree(MST):

After the spanning tree is formed the average weight of all edges W_{avg} and its standard deviation σ is computed. Any edge with a weight $W > W_{avg} + \sigma$ is removed from the tree. This leads to a set of disjoint subtrees $S_T = \{T1, T2, \ldots\}$. Each of the subtrees Ti is treated as a cluster, which has a centre ci. If the number of the subtrees $|S_T| < k$, $k - |S_T|$ additional highest weight edges are removed from the entire edge set of S_T to produce k disjoint subtrees. If $|S_T| > k$, a representative point/centroid is identified for each subtree. Once all the representative points are found, each point in a particular subtree is replaced with the representative point(centroid) of the subtree, thus reducing the number of points in S to $|S_T|$.

A MST is constructed from the centroid of the clusters, and the same tree partitioning process is repeated. When

$|S_T| = k$, the clustering process is considered complete, having produced the required *k* clusters. The formalized minimum spanning tree clustering algorithm is given below.

Algorithm: MSTCLUST (k)

Initialize $n_c \leftarrow 1$

k gives the desired number of clusters.

Let W_e be the weight of edge *e* (as calculated in formula8)

Let σ be the standard deviation of the edge weights

Let $S_T = \Phi$ be the set of disjoint sub trees of the Minimum weight spanning tree

Repeat

Construct a spanning tree from the connected graph G comprising of all patient documents.

Compute the average weight W_{avg} of all the edges

Compute the standard deviation σ of the edges

For each e ϵ MST

If $W_e > W_{avg} + \sigma$

Remove e from MST

$n_c \leftarrow n_c + 1$

$S_T = S_T \cup \{T\}$ //T is the new disjoint subtree

/* If the number of clusters n_c is less than *k*,

remove $n_c - k$ highest weight edges so that $n_c = k$ */

If $n_c < k$

While $n_c \neq k$

Remove the current highest weight edge from MST

$n_c \leftarrow n_c + 1$

$S_T = S_T \cup \{T\}$ //*T* is the new disjoint subtree

Return k clusters

/* If the number of clusters n_c is greater than k */

If $n_c > k$

Compute the centroid *ci* of each Ti € S_T

$S_T = U_{Ti \in ST}\{ci\}$

until $n_c = k$

Return k clusters

3.5 Validation of Clusters

After formation of k-clusters they are validated using Silhouette coefficient which is an efficient validation technique to determine how well each object lies within the cluster [16]. Silhouette coefficient is defined in this section.

For each document *i*, let *a(i)* be the average dissimilarity of *i* with all other data within the same cluster. This paper uses Jacard index ρ(A,B) as the similarity measure and 1- ρ(A,B) as the dissimilarity measure. *a(i)* measures as how well matched *i* is to the cluster it is assigned (the smaller the value, the better the matching). Then find the average dissimilarity of *i* with the data of another single cluster. Repeat this for every cluster of which *i* is not a member. Denote the lowest average dissimilarity to *i* of any such cluster by *b*(*i*). The cluster with this lowest average dissimilarity is said to be the "neighboring cluster" of *i* as it is, aside from the cluster *i* is assigned, the cluster in which *i* fits best. The silhouette coefficient *s(i)* defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$ Equation-9

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

Equation-10

From the above definition it is clear that $-1 \leq s(i) \leq 1$

For *s(i)* to be close to 1 we require *a(i)*<<*b(i).* As *a(i)* is a measure of how dissimilar *i* is to its own cluster, a small value means it is well matched. Furthermore, a large *b(i)* implies that *i* is badly matched to its neighboring cluster. Thus an *s(i)* close to one means that the data is appropriately clustered. If *s(i)* is close to negative one, then by the same logic we see that *i* would be more appropriate if it was clustered in its neighboring cluster.

An *s(i)* near zero means that the data is on the border of two natural clusters. The average *s(i)* of a cluster is a measure of how tightly grouped all the data in the cluster are. Thus the average *s(i)* of the entire dataset is a measure of how appropriately the data has been clustered.

The steps involved in merging of clusters are given below:

Step 1: Select the cluster with least number of patient documents.

Step 2: The documents in the selected cluster are relocated based on the Cluster Merging Criteria, defined in equation-10.

Step 3: The above steps are repeated until no more merging is possible.

4. EXPERIMENTS AND ANALYSIS

The experimental data used in this paper is related to patient wise case study of Liver disease diagnosis and is collected from a reputed private hospital in West Bengal, India whose name could not be disclosed due to confidential terms and conditions. The secondary data is collected from online data sources[17][18]. The training dataset involves different kind of liver disease like cirrhosis, hepato cellular carcinoma(liver cancer), hepatitis, fatty Liver and Wilson's disease etc. The testing data set is used to test the efficiency of the classifier proposed in this paper.

The experimental environment used is: CPU is Intel Pentium Dual Core processor, Memory is 2Gb DDR2 RAM, Windows XP, NetBeans 1.7.0, Java. Table-1 shows specific quantity of samples in each category.

Table-1: Experimental Data Used

Category of Patient data	Quantity of training documents	Quantity of testing documents
Cirrhosis	74	26
Liver Cancer	68	20
Viral Hepatitis	65	18
Fatty Liver	52	15
Wilson's disease	45	12

5. EVALUATION OF CLASSIFIER

During evaluation of text classifiers, both classification accuracy rate and recall rate is considered in the work. After classification, suppose that the number of the documents which are Cj category in fact and also the classifier judge them to Cj category is x; the number of the documents which are not Cj category in fact but the classifier judge them to Cj category is y; the number of the documents which are Cj category in fact but the classifier don't judge them to Cj category is z; the number of the documents which are not Cj category in fact and also the classifier don't judge them to Cj category is d.

Precision is the ratio of the number of documents which judge correctly by classifiers to the number of documents which

classifiers judged to this category, so the precision of Cj is defined as following:

Precision(P) = x/(x+y) Equation-11

Recall rate is the ratio of the number of documents which judge correctly by classifiers to the number of r of documents which are in this category in fact, so the recall rate of Cj is defined as following:

Recall(R) = x/(x+z) Equation-12

We define a Composite Index called

CI= P.R(1+α)/(αP+ R) Equation-13

considering the importance of both recall rate and precision value. Here ($0 \leq \alpha \leq \infty$) is used to show the ratio between precision and recall rate. If $\alpha=0$ then CI=P, if $\alpha=\infty$ then CI=R, and if $\alpha<1$, it emphasizes precision. However in this paper we consider $\alpha=1$ to take into consideration the equal importance of both precision and recall rate. It can be seen from the above table that the traditional KNN algorithm doesn't have satisfactory results and the accuracy is just about 73%, so it is necessary to improve the algorithm in order to enhance the accuracy of classification.

Table-2 Shows the result of KNN classification algorithm.

Category	Precision	Recall	CI
Cirrhosis	0.7627	0.7781	0.77035
Liver Cancer	0.7144	0.7098	0.71204
Hepatitis	0.8013	0.7688	0.78470
Fatty Liver	0.6896	0.6467	0.66745
Wilsondisease	0.7066	0.7130	0.70978

The cluster centers were selected as new training samples and the weight of cluster centers were calculated using (4) mentioned above. So the number of samples reduced drastically which lowered the calculation complexity. Moreover in Table-3 below it can be observed that the accuracy of the proposed classification algorithm improved significantly.

Table-3 Shows the result of SWTCA Classification Algorithm.

Category	Precision	Recall	CI
Cirrhosis(Ci)	0.9024	0.9101	0.9062
Liver Cancer(Lc)	0.8936	0.8798	0.8866
Viral Hepatitis(Vh)	0.9318	.9064	0.9189
Fatty Liver(Fl)	0.8779	0.8493	0.8633
Wilson's disease(Wd)	0.8939	0.9028	0.8983

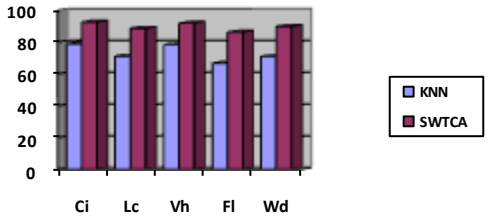

Figure-4: **Comparison of Composite index for KNN and SWTCA**

6. CONCLUSION

This paper proposed an improved similarity weighted text classification algorithm based on clustering, which doesn't use all training samples as traditional KNN algorithm, and it can classify a new document(here patient) automatically to a desired cluster based on similar weight value. This improved algorithm dealt with all training categories by MSTCLUST clustering algorithm for developing k clusters of diseases and finding the cluster center weights which were subsequently used as the new training samples. The clusters were validated using silhouette coefficient. It is observed that an accuracy(CI) of 89% is reached in the proposed algorithm which is much superior to state of art k -NN algorithm for text categorization.

7. FUTURE SCOPE

There are certain issues which will be addressed in future research. More efficient dimension reduction technique will be used for removing less important words. Moreover the clustering algorithm proposed assumes the number of clusters k to be formed is given, future work will consider cases when the no. of clusters is not known.

8. REFERENCES

[1] Jinshu, Su., Zhang, Bofeng., and Xin, Xu (2006). Advances in Machine Learning Based Text Categorization, Journal of Software, vol.17, No.9, pp1848-1859.

[2] Abraham, Ranjit., Jay. B Simha., and Iyengar, S.(2008) Effective Discretization and Hybrid Feature Selection using Naïve Bayesian Classifier for Medical Datamining", International Journal of Computational Intelligence Research, ISSN 0974-1259, vol.4, no.1, pp.974-986.

[3] Wang,Y., and Wang,X.J.(2005). A New Approach to feature selection in Text Classification, Proceedings of 4th International Conference on Machine Learning and Cybernetics, IEEE- 2005, vol.6, pp.3814-3819.

[4] Wang, Yi., Bai,Shi., and Wang, Zhang'ou(2007). A Fast KNN Algorithm applied to Web Text Categorization, Journal of The China Society for Scientific and Technical Information, vol.26, No.1, pp.60-64.

[5] Ying, Li., Zhang, Xiaohui., Huayong, Wang and Chang Guiran(2004). Vector-Combination-Applied KNN Method for Chinese Text Categorization, Mini-Micro Systems, vol.25, no.6, pp.993-996.

[6] Lee,L.W., and Chen, S.M.,(2006). New Methods for Text Categorization based on a new feature selection method and new Similarity measure between Documents, IEA, France.

[7] Chapman, W.W., Dowling, J.N., and Wagner, M.(2004) Fever detection from free-text clinical records for biosurveillance, Journal of Biomedical Informatics, volume 37, Issue 2, pp.120-127.

[8] Mamlin, B.W., Heinze,D.T., and McDonald, C.J(2003) Automated Extraction and Normalization of Findings from Cancer-Related Free-Text Radiology Reports, Proceedings of the AMIA , pp 420-424.

[9] Mukherjea,S., Bamba,B., Kankar,P(2005) Information Retrieval and Knowledge Discovery Utilizing a BioMedical Patent Semantic Web", IEEE Transactions on Knowledge and Data Engineering, Volume 17, Issue 8, pp.1099 – 1110.

[10] Pakhomov,S., Hanson,P.L., Bjornsen,S,Smith,S(2008) Automatic Classification of Foot Examination Findings Using Clinical Notes and Machine Learning", Journal of American Medical Informatics Association,15(2):198202.doi:10.1197/jamia.M2585.

[11] Peter,John,S.,(2010) Minimum Spanning Tree-based Structural Similarity Clustering for Image Mining with Local Region Outliers, International Journal of Computer Applications (0975 – 8887) Volume 8, no.6.

[12] Sebastiani, Fabrizio(2002). Machine learning in text categorization, ACM Computer Survey, vol.34, no.1pp.1- 47.

[13] Ranjani,R., Anitha,S., Elavarasi and Akilandeswari,J.(2012) Categorical Data Clustering using Cosine based similarity for Enhancing the Accuracy of Squeezer Algorithm, International Journal of Computer Applications, 45(20), pp. 41-45, Published by Foundation of Computer Science, New York.

[14] Asano,T., Bhattacharya,B., Keil,M., and Yao,F(1988) Clustering algorithms based on minimum and maximum spanning trees. In Proceedings of the 4th Annual.Symposium on Computational Geometry, pp.252–257.

[15] Satu Elisa Schaeffer (2007). Survey on Graph Clustering, Elsevier Computer Science Review,pp.27-64, doi:10.1016/j.cosrev.2007.05.001.

[16] Aranganayagi,S, Thangavel, K(2007). Clustering Categorical Data using Silhouette coefficient as a relocating measure, In Proceedings of International Conference on Computational Intelligence and Multimedia applications,0-7695-3050-8/07, DOI 10.1109/ICCIMA.2007.328

[17] www.hepatitis.va.gov/provider/cases/index.asp. The site provides patient cases which were presented at the Veteran Affairs Advanced Liver Disease Resource Training Programs and illustrate a number of areas of liver disease that are actively being researched.

[18] www.livestrong.com/article/41489-liver-surgery,Complications. The site livestrong.com highlights the complications that arise from liver surgeries like bile duct problems, rejection on transplant, infections and thus emphasizes that patients for liver surgeries should be carefully chosen.

APPENDIX-A

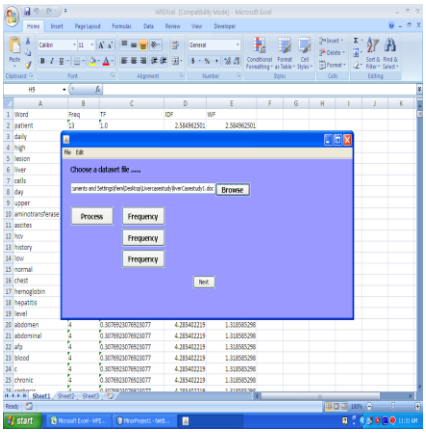

Figure-5: A snap shot of the run time environment used during weight factor calculation and the interface used to take input of patient cases in the form of text files.

APPENDIX-B

Table-4: Shows the relationship between top ten words and their frequencies retrieved from patient cases.

Rank	Word	Frequency
1	liver	188
2	Hcv	171
3	Hepatic	168
4	Lesion	147
5	Chronic	144
6	Abdomen	136
7	Test	133
8	Ultrasound	117
9	Biopsy	106
10	Infection	103