INTERNATIONAL JOURNAL OF SOFT COMPUTING

MEDWELL PUBLICATIONS

Pragmatic Assessment of Occurrence of Brain Cancer with Incidence Levels using Collaborative Big Data Mining Techniques

Syed Rizwan, Venu Madhav Kuthadi and Rajalakshmi Selvaraj
Department of CS and IS, Botswana International University of Science and Technology, Botswana, Southern Africa

Key words: Brain cancer, incidence levels, big data mining, stages of brain cancer, k-means clustering, big data analytics

Corresponding Author:
Syed Rizwan
Department of CS and IS, Botswana International University of Science and Technology, Botswana, Southern Africa

Page No.: 61-67
Volume: 14, Issue 3, 2019
ISSN: 1816-9503
International Journal of Soft Computing

Abstract: The major objective of this research is to identify the presence of brain cancer along with the incidence levels of beginning stage to advanced stage using collaborative analysis of big data and data mining techniques. The dataset collected from secondary sources had few errors and rectified using preprocessing techniques in MATLAB. Further, the testing dataset is processed with k-means algorithm to form cluster analysis and identify the presence of brain cancer in three levels of well, fair and poor levels using degree of difference between the normal and cancer cells in brain. The algorithm is modified according to the needs of the medical analysis of the current dataset. The results indicates the presence of brain cancer in various three levels under cluster values of initial stage (54%), Curable stage (38%) and incurable stage (8%), respectively. The accuracy of prediction is 93.4% and the error identification is 9.3% whereas the sensitivity and specificity accounts to 0.8 and 0.7, respectively. Hence, further analysis is conducted in tableau big data tool and the sheets with story boards are formed. This research indicates the occurrence of brain cancer is influenced by gender and age factors along with regular activities and streams. Thus, brain cancer is considered as one of the challenging prediction as the cell contains mixed patterns with variations according to gender and age of human beings.

INTRODUCTION

The insightful fight against cancer is the order of the medical world in recent times. Various medical and research organization are involved in eradicating the deadliest disease cancer which is considered as one of the serious threat for humanity in recent years. Earlier, cancer is a rare disease which affects the people who had bad ideals in life. In recent cases it is extremely different as even normal persons with no bad habits, children and women are getting affected due to modernized way of living and culture. It is highly prevalent in almost all countries and hence the medical world is in search of good research ideas that with kindle the hopes of

identifying the presence of cancer before it actually spreads to maximum level. Brain cancer is one type of cancer that could be even worse compared to the damage that it can cause in a human being. Since, cancer is affected over a million of people around the world, it is very hard to diagnose the reasons for the incidence of the disease. Hence, collaboration of data mining and big data could be a better solution in identifying the levels of brain cancer at the initial stage itself. Data mining is an art of extracting knowledge by mining patterns of algorithms with the existing dataset whereas big data is the recent innovation to identify complicated patterns and promote successful outcomes from large and complex datasets. Cancer is a stage where there is abnormal growth of cells in the human body. Hence, through cell structure and genomics, it is easy to identify the cancer at any stage. The objective of this research is to identify and predict the presence of brain cancer in various levels of incidence and its outcome analysis with perspective to age and gender in human beings. Thus, to conduct this research, various reviews are identified and analyzed from various authors.

Literature review: Data mining is highly predictive in case of medical diagnostics and hence collaborated with big data which can handle huge medical dataset compared to other technologies. Still much of the research ideas are in the premature stage as it involves huge amount of data to be analyzed and presented. Few of the research studies are presented in the review of brain cancer with big data and data mining.

Kiranmayee *et al*. (2017) formulated a hybrid algorithm in data mining to predict the occurrence of brain tumor using normal clinical brain dataset. The brain tumor is a form of intra cranial neoplasm that has irregular growth in human brain. This kind of tumor follows up with sever damages in frontal, temporal and parietal part of the brain. The author used machine learning algorithms with clustering, classification and association methods to interpret, predict and analyze the results.

Ramani and Sivaselvi (2017) endeavored to conduct an analysis on the performace of various supervised algorithms in data mining to classify brain Magnetic Resonance images. The images collected are preprocessed to avoid any damages and then involved with feature selection technique like analysis, fisher filtering and relief feature selection to determine best features. The author used Naive Bayesian, support vector machine, random tree and C4.5 to identify the abnormal images present in brain. The best accuracy is identified with SVM algorithm with 71.33% whereas Random tree achieved 82% with run filtered features only.

Ahmad and Aziz (2017) conducted a research to detect brain cancer in obese and non-obese patients through classification technique in data mining. This research being a broader one for both patients and doctors involves huge and noisy dataset. The research used c5.0 algorithms with preprocessing of dataset and predicted the obese and non-obese patients to identify the presence of brain tumor.

Yin *et al*. (2017) focused on the relationship between the structure of brain and brain metastases of occurrence was analyzed with patients. It used 20 patients for developing the initial model and sixty nine data for validation and verification of the model. The MRI images identified 116 brain regions from the GM volume. The LOOCV test achieved highly correlated values for 20 datasets with r =0.834. The other results also showed over 74% accuracy based on the clinical dataset for brain.

Jalali *et al*. (2017) created a framework with data mining to detect breast cancer in Iran hospitals in which association rules for reducing the size of datasets are applied with classifiers for predicting the cancer. A k-fold cross validation is applied for identifying the performance of six breast classifiers using SVM algorithm. The prediction accuracy achieved is 93% for detecting the presence of cancer.

Gupta and Ahmad (2017) represented a combination of SVM and Fuzzy algorithm for brain cancer detection. The fuzzy is specifically fuzzy c-means which used techniques like spatial domain method and frequency domain method for prediction. The gray level co-occurrence matrix used for texture features from brain pictures to identify the tumour from non-tumour cells.

Horvat and Kochanek (2017) conducted a big data research on well conducted tests with negative impact on adults and children with high possibility of getting cancer. The test was conducted in three levels with management components, work to be done and to be considered respectively. The research showed clear indication of utilizing the analytics for cancer like diseases in medical diagnosis.

Manogaran *et al*. (2017) applied Bayesian Hidden Markov Model (HMM) with Gaussian mixture to create a model to predict and identify the DNA copy number change that could be applied across the human genome. The research conducted pruning of data using binary segmentation and results are proposed to demonstrate the effective accuracy of the brain cancer prediction.

Among all the disease predictions applied in the various researches conducted earlier, none of the cases utilized the numeric dataset for predicting the brain cancer occurrence and to identify the presence of incidence levels of brain cancer in human body.

MATERIALS AND METHODS

The research that is previously conducted comprised of MRI images with segmentation analysis using various data mining algorithms to identify the presence or absence of brain tumor. Hence, the feasibility study is conducted and the dataset collection is proposed. The overall methodology involves 5 major stages as indicated in the Fig. 1:

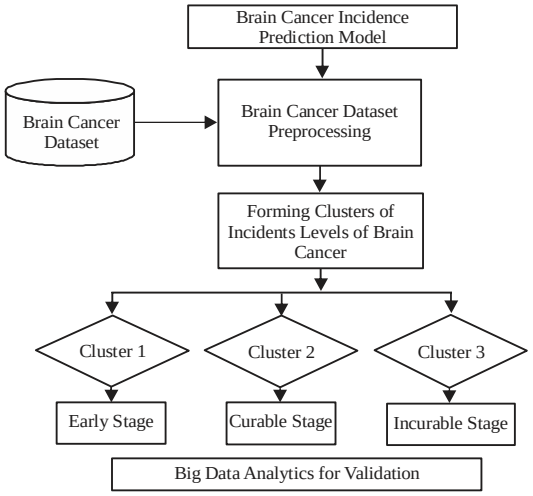

Fig. 1: Framework of brain cancer incidence prediction

- Dataset collection
- Preprocessing of dataset
- Prediction using k-means algorithm
- Formation of clusters for brain cancer incidence levels
- Conducting analytics using bi data tools

The dataset for brain cancer is collected initially and then processed with data mining algorithms to form refined dataset. The dataset is then applied with correctness and validation test to generate the testing dataset. The testing dataset is applied with k-means algorithm to form clusters of data. The clusters indicates the incidence of brain cancer in three levels as early, curable and incurable stage respectively. Finally, the dataset is applied with big data analytics to predict the incidence levels in graphical format. This is one form of verifying the results thereof. The methodology is a new dimension for predicting brain cancer as the existing system in many cases shows only image based processing and diagnostics methods. However, numerical dataset is a new innovation in recent times where many research organizations prefer to utilize the same for improving their accuracy in predictions comparing to image processing as it could provide less reliable output.

Preprocessing of dataset: The dataset for the research is collected as a secondary resource through UCI repository where the dataset contains 339 instances and 18 attributes including one class attribute for prediction. All the attributes have been entered in the database as numerical data that is processed from images and then classified for correct input. The class attribute indicates the type of region affected with cancer which includes lung, head and neck, thyroid, stomach, colon, rectum, pancreas, brain, liver, kidney, testis, prostate, breast, etc. The other common attributes used in the dataset are in Table 1.

Table 1: List of attributes and range values

Attribute	Range values
Age	<30, 30-59, ≥60
Sex	Male, female
Histologic type	Epidermoid, adeeno, anaplastic
Degree-of-diffe	Well, fairly, poorly
Bone	Yes, no
Bone-marrow	Yes, no
Lung	Yes, no
Pleura	Yes, no
Peritoneum	Yes, no
Liver	Yes, no
Brain	Yes, no
Skin	Yes, no
Neck	Yes, no
Supraclavicular	Yes, no
Axillar	Yes, no
Mediastinum	Yes, no
abdominal	Yes, no

Initially the dataset collected is found to contain few errors like missing data, irrelevant data, non-numeric data, etc. Hence, the dataset is preprocessed using data mining algorithms (Fig. 2).

Hence, after preprocessing, the accuracy of the complete dataset is found to be 93% which is considered to be good for predicting the results with data mining algorithms. Thus, after pre-processing, clusters are formed.

Formation of clusters: The preprocessed dataset contains 18 attributes that includes 17 normal attributes and one class attribute. Hence, after analysis 6 attributes are selected for forming clusters using k-means algorithm. This is finalized using the MATLAB graphical assessment of dataset (Fig. 3). They are:

- Class attribute: to indicate the region affected with cancer
- Age attribute: to indicate age range of patient
- Sex attribute: to indicate gender of patient
- Histologic-type attribute: to indicate type of cancer
- Degree-of-difference: to indicate incidence levels
- Brain: to indicate presence or absence of brain cancer

The k-means is an algorithm in data mining that is capable of forming clusters to segregate and segment the cells affected under different console. Thus, in brain cancer, k-means is applied and the incidence levels are identified using MATLAB using centroid values. The centroid values differentiate the incidence level of brain cancer in three types:

- Early stage: which can be cured naturally
- Curable stage: which can be cured after treatment and medications
- Incurable stage: which cannot be cured and treated. But through medications, it can be maintained

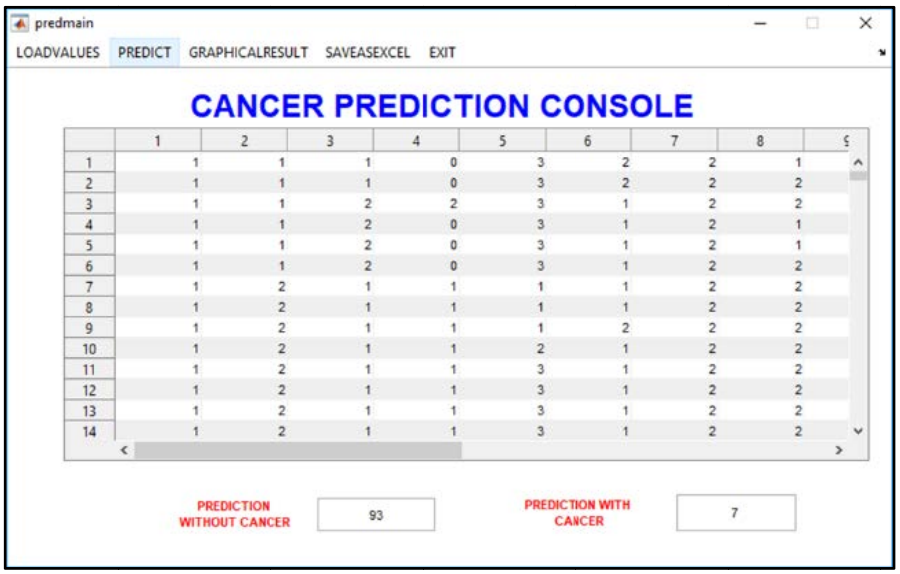

Fig. 2: Preprocessing of brain cancer dataset

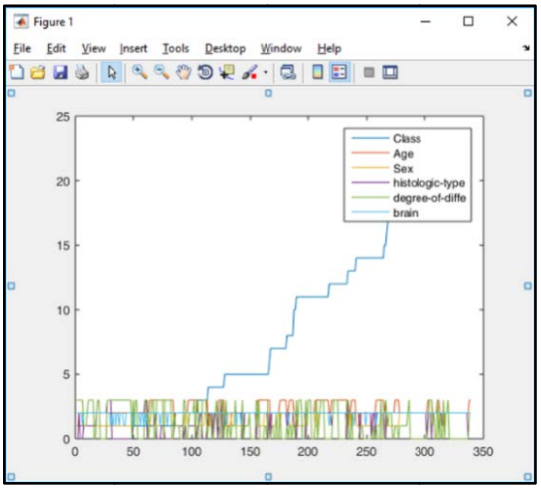

Fig. 3: Selection of Attributes for prediction

Based on the created dataset, the implementation is carried out in MATLAB 2015 in the presence of data mining algorithms applied on the numerical brain dataset.

Implementation and predictions: The implementation of the k-means algorithm begins with the formulation of testing dataset that should be applied with the user interface design of the research. The ultimate goal of the implementation is to predict the presence of brain cancer and also to check the incidence levels of the algorithm under three stages of brain cancer. The overall mathematical formulation applied with k-means is:

$$\text{Cluster} = \sum_{0}^{330} x[i] = \text{Centriod}$$

Algorithm BCPKM

Variables:

clust1 = 0; clust2 = 0; clust3 = 0; canc = 0; ncanc = 0;

sens = 0; spec = 0; accur = 0; erper = 0;

Begin

for i=1:330

if brain(i) == 1//indicates brain cancer

Increment Canc by 1;

if Histologic-type Attribute (i) == 1

Increment cluster1 by 1;

else if Histologic-type Attribute (i) == 2

Increment cluster2 by 1;

else if Histologic-type Attribute (i) == 3

Increment cluster3 by 1;

Else ncanc = ncanc + 1;

if degree_of_difference (i) == 1

Increment cluster1 by 1;

else if degree_of_difference (i) == 2

Increment cluster2 by 1;

else if degree_of_difference (i) == 3

Increment cluster3 by 1;

else

Increment ncanc by 1;

Compute sens = (clust3)./(ncanc);

Compute spec = (clust2 + clust1)./ (ncanc);

Compute accur = log(canc) / log(clust1 + clust2 + clust3) * 100;

Compute erper = 1 - (log(ncanc ./ canc) / log(clust1 + clust2 + clust3)) * 100;

End.

Fig. 4: Algorithm brain cancer prediction k-means (BCPKM)

The algorithm is applied in the Matlab 2015 with GUI design as follows Fig. 4 and 5.

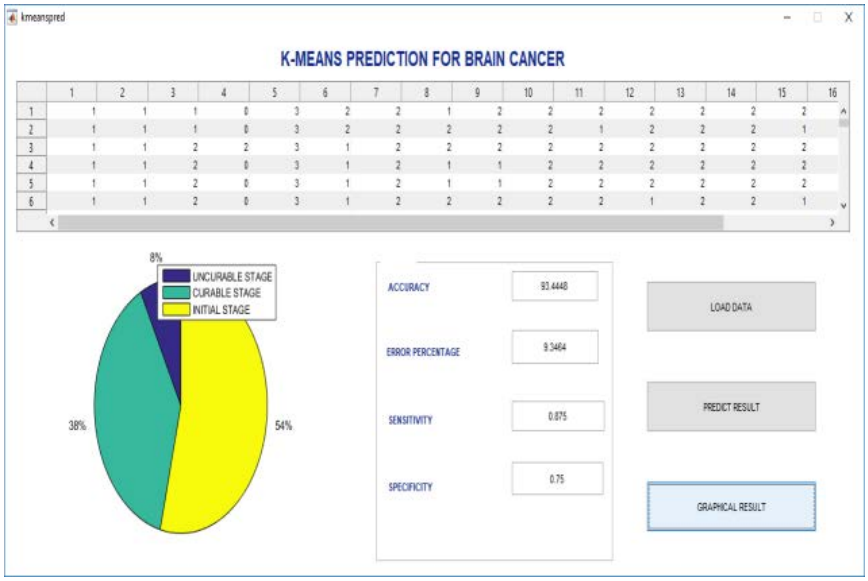

Fig. 5: k-means result with GUI implementation

Table 2: Accuracy table of k-means prediction

Parameters	Values
Accuracy	93.4448
Error Percentage	9.3464
Sensitivity	0.875
Specificity	0.75

Table 3: Brain cancer incidence levels

Levels	Percentage
Early stage	54
Curable stage	38
Incurable stage	8

Initially the testing dataset created after preprocessing is loaded into the MATLAB table and then applied with the k-means algorithm to predict the incidence of brain cancer in human beings. Thus after the algorithm is implemented using the design in MATLAB, the predictions are comparatively good with other algorithms (Table 2 and 3).

Thus, the predictions of the brain cancer is possible with the implications of data mining algorithms. However, it is also important to identify the relevance of data related to big data analytics. Hence, Tableau is utilized for graphical predictions of the brain cancer.

Big data analytics: Big data is a technology that is capable of handling huge data that could be very hard to manage by data mining which can handle only limited data. As a follow up of the implementation and the data mining prediction for incidence of brain cancer, the tableau also made graphical predictions for the testing dataset after loading into it.

The analytics indicated that the average persons affected with brain cancer accounted to 60-80% of the total cancer affected persons. Hence, after analytics, the results are discussed on the overall research conducted to predict the occurrence of brain cancer and also on the incidence levels of brain cancer.

RESULTS AND DISCUSSION

The research is an attempt to understand and propose a methodology that can predict the dataset with maximum accuracy using numerical data rather than image scans and image processing methods. The first method preprocessing removed all the irrelevant data from the dataset, especially, the symbolic data and non-numeric data. The results confirmed that a dataset with 93% accuracy is considered as sample and assessed using k-means algorithm (Fig. 6).

The application of k-means clustering algorithm to predict brain cancer presence and also its incidence levels also identified the accuracy of 92% which is much appealing compared to other algorithms that existed earlier (Fig. 7).

After the k-means prediction, the big data analysis is also carried out and the overall storyboard indicated the following results (Fig. 8 and 9).

The result indicates that out of 330 respondents 195 members are affected with brain cancer. It is a heavy result compared with other cancer results in the class attribute. Thus, these results predicts that brain cancer can be assessed using numerical datasets.

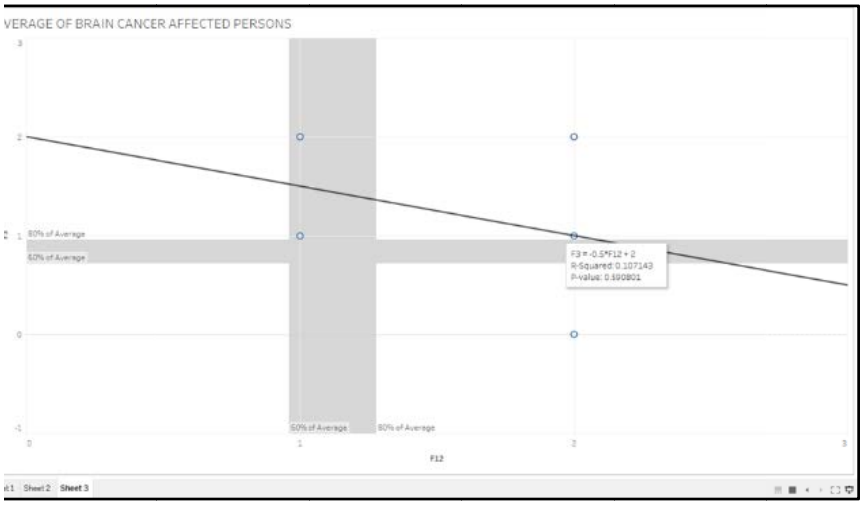

Fig. 6: Average levels of brain cancer using sheet analysis in tableau

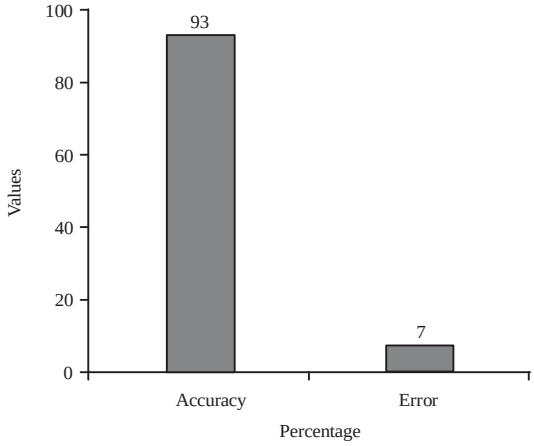

Fig. 7: Accuracy of testing dataset after preprocessing

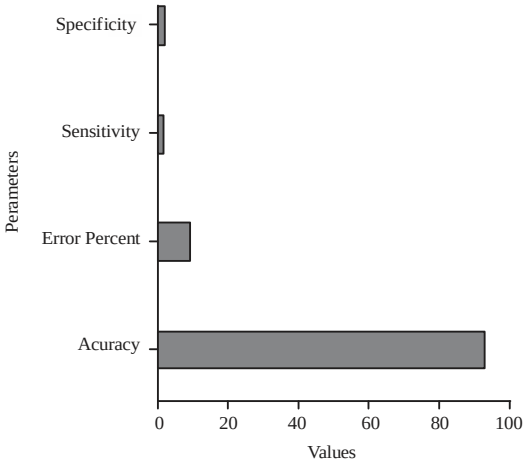

Fig. 8: Accuracy of k-means prediction for brain cancer incidence levels

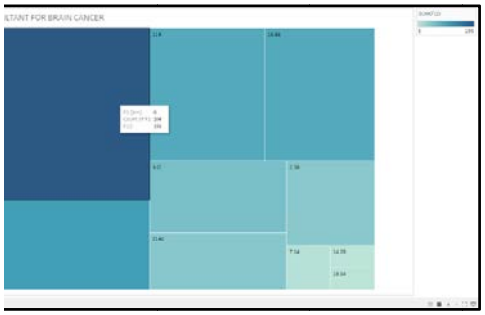

Fig. 9: Overall outcome as storyboard for brain cancer

CONCLUSION

The research presented a new dimension of brain cancer evaluation apart from normal levels like image processing techniques. Thus the overall results indicated that the collaboration of data mining with big data could bring much bigger results compared to individual assessment of records. The accuracy and the results are convincing and kindles for further research in the future. This research could be extended to sensor based IOT predictions that involve complex algorithms and could be implemented in the hospitals. This research has a cause and implemented for the benefit of the society to eradicate brain cancer and pave way for a happy and productive life of people in the world.

REFERENCES

Ahmad, M. and A. Aziz, 2017. Early detection of brain cancer in obese and non-obese patients by using data mining techniques. Indian J. Sci. Technol., Vol. 10, No. 24.

Gupta, D. and M. Ahmad, 2017. A hybrid technique based on fuzzy methods and support vector machine for prediction of brain tumor. Int. J. Comput. Sci. Eng., 9: 517-521.

Horvat, C.M. and P.M. Kochanek, 2017. Big data not yet big enough to determine the influence of intracranial pressure monitoring on outcome in children with severe traumatic brain injury. JAMA. Pediatr., 171: 942-943.

Jalali, S.M.J., S. Moro, M.R. Mahmoudi, K.A. Ghaffary and A. Alidoostan, 2017. A comparative analysis of classifiers in cancer prediction using multiple data mining techniques. Int. J. Bus. Intell. Syst. Eng., 1: 166-178.

Kiranmayee, B.V., T.V. Rajinikanth and S. Nagini, 2017. Effective analysis of brain tumor using hybrid data mining techniques. Int. J. Adv. Res. Comput. Sci., 8: 286-293.

Ramani, R.G. and K. Sivaselvi, 2017. Classification of pathological magnetic resonance images of brain using data mining techniques. Proceedings of the 2017 2nd International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), February 3-4, 2017, IEEE, Tindivanam, India, pp: 77-82.

Yin, G., C. Li, H. Chen, Y. Luo, L.C. Orlandini, P. Wang and J. Lang, 2017. Predicting brain metastases for non-small cell lung cancer based on magnetic resonance imaging. Clin. Exp. Metastasis, 34: 115-124.