Advanced Computing: An International Journal (ACIJ), Vol.2, No.5, September 2011

PRE-PROCESSING OF MEDICAL DOCUMENTS AND REDUCING DIMENSIONALITY

S.Sagar Imambi *, T.Sudha**

*Department of Computer Science, TJPS college ,Guntur, India
simambi@gmail.com
**Prof & Head , Department of Computer Science Vikram Simhapuri University

ABSTRACT

The exponential growth of online repositories in medical science has led to the development of various text mining tool . Theses tools assist the users in analyzing text data stored in the online repositories like Pubmed and Medline. The pubmed repositories are growing at the rate of 500000 articles per year. Classification of Medline documents becomes very complex due to high dimensionality of feature space. In this study we discussed how dimensionality is reduced. We study and compared various dimensionality reduction techniques at the pre-processing stage. We introduce a novel feature weighting scheme 'GRW ' and proved that this schema improves classification accuracy. Our experimental results indicate that existing feature weighting methods has less accuracy rate when compared to GRW schema and tested on medical data set

KEYWORDS

Medline Documents, Pre processing ,Text mining .

1. INTRODUCTION

Medline and Pubmed repositories are rich in medical literature, supported by National library of Medicine. The number of available articles currently in PubMed are 15,000,000, and this number grows by the hour. It makes very difficult to find the documents relevant to a specific need.Automatic extraction of useful information from these online sources remains a challenge because these documents are unstructured and expressed in a natural language form i.e in text format. The abundance of these data and literature produces a major bottleneck for Interpreting.. The ability to rapidly survey this literature is therefore a necessary step. Automated text mining integrate information gathered from multiple documents and helps in analyzing the documents.

Text mining becomes very supportive tool for Medline and Pubmed repositories classification Text mining tools are first developed in order to facilitate the automated searching of digital library material by users [2]. Due to advent of powerful computing facilities and widespread of www, text mining becomes a new and exiting research area. Text mining applies techniques like Data mining, knowledge management and information retrieval and NLP to solve information overload problem. Data mining is used for medical decision making [9].

DOI : 10.5121/acij.2011.2502

2. TEXT MINING

Main phases of text mining are 1. Text gathering 2) Text preprocessing 3) Data analysis 4) Visualization

Text gathering includes collection of raw documents. They are unstructured data .Preprocessing phase starts with tokenization. Tokenization is division of a document into terms. This process also referred as feature generation. It removes stop words and apply stemming algorithm to represent the terms in stemmed form. The stemmed words are input for feature selection and reduction. After features are reduced , the documents are represented in vector format and are ready for analysis Ronen Feldman,(2007). In our previous work we show text classification is used to predict the risk factors of Diabetic retinopathy S.Sagar Imambi(2010)

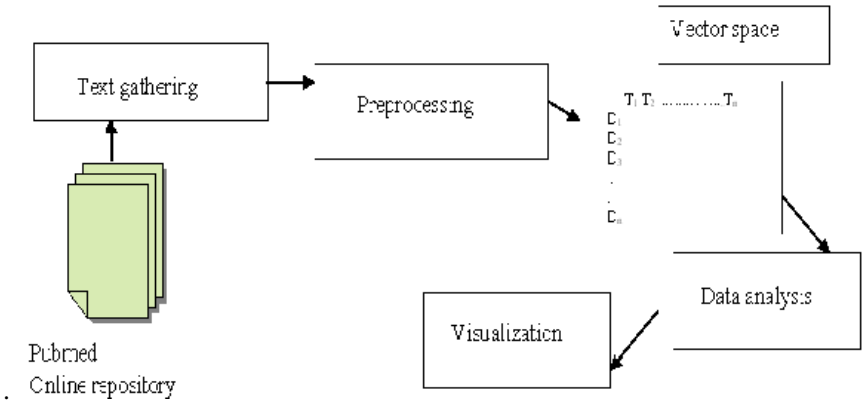

Fig 1: Text mining architecture

In Data analysis stage several data mining tools like Classification, clustering and association rules are applied on vector space model of data. The analysed reports are visualized in the last phase in terms of graphs.

Pubmed articles are indexed by MESH terms. Mesh heading and sub heading are powerful tool indexing tools. As Pubmed repositories are growing at the rate of 5,00,000 articles per year manual indexing becomes very difficult process[1]. We require special text mining techniques like text categorization or text classification. The fig 2 shows the Mesh index for Diabetes complications. We collected Pubmed abstracts , which are published between 2000 and 2010. We try to improve the Medline classification process by reducing the number of keywords. In this paper we focused on preprocessing techniques of Medline documents . We study and compared various dimensionality reduction technique at the preprocessing stage.

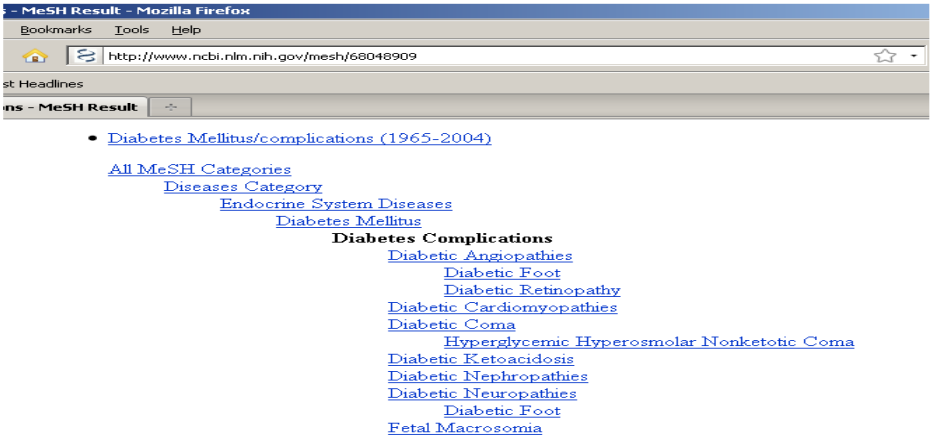

Fig 2. Diabetes Mellitus Complication ,Source: Http://www.ncbi.nim.nih.gov

2.1 Preprocessing :

The raw data in the form of text files is collected from online repository pubmed . The data is converted in to xml files and stored in Database. The architecture of Preprocessing stage is shown in the fig 3. The preprocessing usually includes converting xml documents into text document, removing stop word, performing word stemming. Stop words are very frequently used common words like 'and' are' 'this' e.t.c. They are not useful in classification of documents. So they must be removed. Word stemming removes suffixes and generate the stemmed words ex. Retrieval becomes retrie. We used Porter Stemmer algorithms for word stemming.

2.2 Feature generation:

Extracting relevant feature from the text files is called feature generation. The main goal of feature generation is to transform a document in to a list of relevant features or keywords. Feature generation methods are classified into two main classes. Filter methods and wrapper methods. Filter methods use an evolution function that depends on data and is independent of inductive algorithm.(Sima C et al 2006) Wrapper methods use inductive algorithms to estimate the value of given subset. The inductive algorithm induces a classifier which is useful in classifying future set. The classifier is mapping from the space of feature values to the set of class values.

2.3 Feature Selection:

The generated features are assigned weights using various weighting techniques. The feature selection algorithm conducts a search for best subset using valuation algorithm. The valuation algorithm is run on the dataset usually portioned into internal training and test set with different set of features removed from the data. The feature subset with the highest evolution is chosen as the final subset on which to run the induction algorithm

2.4 Indexing:

After selected keyword(features) the terms are indexed and the whole document set is represented in vector space model. Vector space model uses term-document matrix notation. The representation in this model is as follows

$$D= \{d1,d2,,d3,d4\ldots\ldots dn\}$$

d1={t1,t2,t3,t4, …….tm}

where D indicates the total document set with n elements and d1 is set of m terms. The matrix element [i,j] identifies term tj in ith document. The dimension of vector space is 'm' . The dimension is equal to no of the terms (reduced by feature selection) in the documents.

1.http://www.ncbi.nim.nih.gov

3. STATEMENT OF THE PROBLEM:

Set of documents are extracted from pubmed giving the related query i.e Type 2 diabetes complications. We would like to automatically reduce the dimensionality (features) of the data set by applying pre processing and Feature selection algorithms. The main objective of this study is to improve the accuracy of classification of Medline documents by removing the irrelevant , noisy features and compare the precision and recall of various Feature selection methods. The general notations used are

D={d 1,d 2, ... , d n}: the training document collection

C ={c1,c2, ... , cm }: the set of possible categories to be assigned to the documents.

T ={t 1,t 2, ... , t m}: the set of terms appearing in the documents.

wij: the weight of thej th term of the i th document.

N = Total number of documents.

IDFj =Global weight of the term j.

DFj = The document frequency of the term j.

NCi = Number of documents in class i.

NT_{ij} = The total of documents that contain $term_i$ and belongs to class j.

4. SURVEY OF LITERATURE:

Filter approaches evaluate the relevance of features using data set. They are not depend on the classification algorithms. Some of popular Filter Feature selection methods are TFIDF, Information gain, Chi square, gain ration, term strength, Mutual Information, CFS e.tc.

TFIDF is one of the first weighting schemas and used to select the features. TF is the term frequency and IDF is inverse Document frequency and is calculated by TF* log (N/DF_i) where

DF is document frequency of term i and N is total number of documents . It becomes base line for many term weighting studies.(K.Peripinani 2001, Robertson et al 2004)

Information Gain measures entropy of features. Entropy measures the no of bits of information obtained for class prediction by knowing the presence or absence of term in document. From the training data IG is calculated for each term and the terms whose IG is less than the threshold are selected.(Jnovovicova et al 2004). Uguz H(2011) shows that combined features of PCA and IG improves the accuracy of the classification of brain disease .

Chi square evolution is generally used in statistical analysis and measures the lack of independence between term and class. The difference between IG and Chi square is , chi square uses normalized values and is not reliable to low frequency terms.(Yang. Y et al 1997).

ChiWSS is variation of chi square, proposed by Ranjit Abraham et al (2007). This method improves the classification accuracy of Navie Bayes with respect to medical dataset. They used wrapper approach to reduce the dimensionality by eliminating the irrelevant features using chi square statistics. The Feature selection performance is tested with SVM and logistic regression models.

Term strength calculates only the number of documents the term contains and there are several variations of this method like TFIDF and so on. Ng. et al(2006) proposed FS method called WLLR (Weighted log like hood ratio) . He achieves the accuracy of 87% with the test data .WLLR is dealing with terms with high category ratio and high document frequency. The formula for calculating WLL(t,Ci)= P(t/Ci) * Log(P(t/Ci)/P(t/Ci') .

Relief [18] and its multiclass extension ReliefF are supervised feature weighting algorithms based local weights.

Chi-square [19] is used to assess two types of comparison: tests of goodness of t and tests of independence. In feature selection it is used as a test of independence to assess whether the class label is independent of a particular feature. Forman (2003) presented an empirical comparison of twelve feature selection methods. Results revealed the surprising performance of a new feature selectionmetric, 'Bi-Normal Separation' (BNS).

Guyon and Elisseeff (2003) , Ann Li (2009) proposed WFO (weighted frequency and ODDS) This method is robust when the set contain large number of features. The parameter is used for tuning the weight between frequency and odds .

5. PROPOSED ARCHITECTURE FOR PREPROCESSING MEDLINE DOCUMENTS:

The proposed architecture includes preprocessing layer and feature selection layer Preprocessing includes tokenizing, stemming, stop words removal etc. we used the list of 1200 stop words. The feature generation extract unique terms from documents and weighted by using the novel global relevant weighting schema The proposed schema is variation of global weight schema IDF. GRW(t) is calculated by using the below formula**.**

$$GRW(tci)= TFIDFj * P(Tij)/P(Ci)$$

Where TFIDFj is the tfidf value of jth term, P(Tij) is probability of term 'j' belongs to class 'i' and P(Ci) is the probability of documents that belongs to class 'i' . Feature selection strategy is applied to select GRW of the term by using selection criteria

$$GRW(t)= \max\{ GRW(t)/Ci\}.$$

For example ,
If GRW(t1,c1)=0.25,
GRW(t1,c2)=0.28 and
GRW(t1,c3) = 0.4 then term t1 is selected for class 3.

All the terms with high relevance are selected from each class and indexed in the Indexing phase. Threshold 'th' is used to select the terms from each class. As the dimension is reduced ,now the documents are represented in document column represents the term and the value in vector is corresponding relevant weight of term.

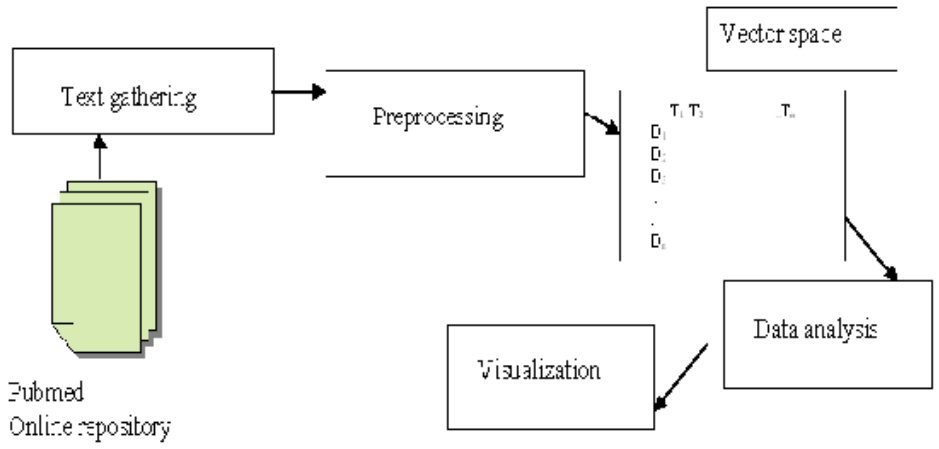

Fig 3 Preprocessing architecture

6. EXPERIMENTAL STUDY:

6.1 Dataset:

We collected 5610 document instances from pubmed online repository which are deposited between 2000-2010. The dataset includes documents related to Diabetes Mellitus Complications. I choose only cardiomypothesis, neuropathy and retinopathy according to Mesh tree structure fig 2.2. After removing nosiy documents the data set size becomes 5460.The data set is labeled as ('cardio, 'neuro','retina').The no of documents under each category are cardio 680, neuro 3030 and retina 1750.

6.2 Experimental Setup: By using Matlab we developed software for generating features and use the novel approach for feature selection. As Matlab is very flexible in vector processing, we developed a program that generates vector space model of documents with various weighing schemas. The weka software is used to test the accuracy of various existing feature selection techniques available in weka and compared with our new method.

6.3 Experimental result:

The Feature evolution measures effectiveness of learning algorithm. The feature set is evaluated based on the performance of learning set Accuracy , Precision and Recall are the best measures in this field. Accuracy is the ratio between total number of documents and the no of the documents correctly classified. Precision is the percentage of the documents that are correctly classified . Recall is the percentage of total documents that are correctly classified.

The formulas for Precision is TP/(TP+FP) and recall= TP/(TP+FN) and accuracy is TP+TN /TP+FP+TN+FN These terms are obtained from the confusion matrix.

	TRUE	FALSE
TRUE	TP	FP
FALSE	FN	TN

Table 1.confusion matrix

While measuring Precision and recall values are calculated for each classifier and tabulated in the table 2,3. Very popular classifiers like BayesNet, NaiveBayes ,Decision tree (Cart) , Decision table are used for testing accuracy.

The proposed 'GRW' Selection method is compared with TFIDF, CFS, Gain ratio, Chi square , and Filtered subset . The results are tabulated in the below tables. Table 2 shows the accuracy with the famous classifiers.

FS method	Cart	Decision-table	Bayesnet	Bayes
GRW	99.2674	99.084%	100%	70.8794%
TF	65.9341%	65.9341%	65.2015 %	58.2484
CFS	64.652	63.7363	65.567%	66.8498 %
Gain raio		63.7363	65.9341%	61.1722 %
Chisquare	64.8352 %	63.7363 %	65.9341%	57.6929
Filteredsubset	64.28.57	64.2857%	65.5678 %	65.5678 %

Table 2 Accuraccy of various Feature selection methods

FSmethod	Cart	Decision-table	Bayesnet	Bayes
GRW	0.993	0.991	1	0.753
TF	0.62	0.615	0.616	0.594
CFS	.612	0.566	0.586	.661
Gain raio	0.576	0.576	0.58	0.618
Chisquare	0.565	0.576	0.58	0.587
Filteredsubset	0.566	0.566	.586	0.586

Table 3 Precision of various Feature selection methods

FSmethod	Cart	Decision-table	Bayesnet	Bayes
GRW	0.993	.991	1	0.709
TF	0.659	0.659	0.652	0.582
CFS	0.647	0.643	0.656	0.668
Gain raio	0.637	0.637	0.658	0.612
Chisquare	0.648	0.637	0.659	0.577
Filteredsubset	0.643	0.643	0.656	0.656

Table 4.Recall of various Feature selection methods:

The fig 4 show the result of Cart classification with GRW feature selection using WEKA software. It give 99.267% of accuracy. The fig 5 represents the accuracy of the feature selection methods with 4 classifiers. X axis represents the feature selection methods. '1' is our proposed method 'GRW'. This graph indicates that accuracy is very high for GRW and GRW schema is best suited for Medical literature classification.

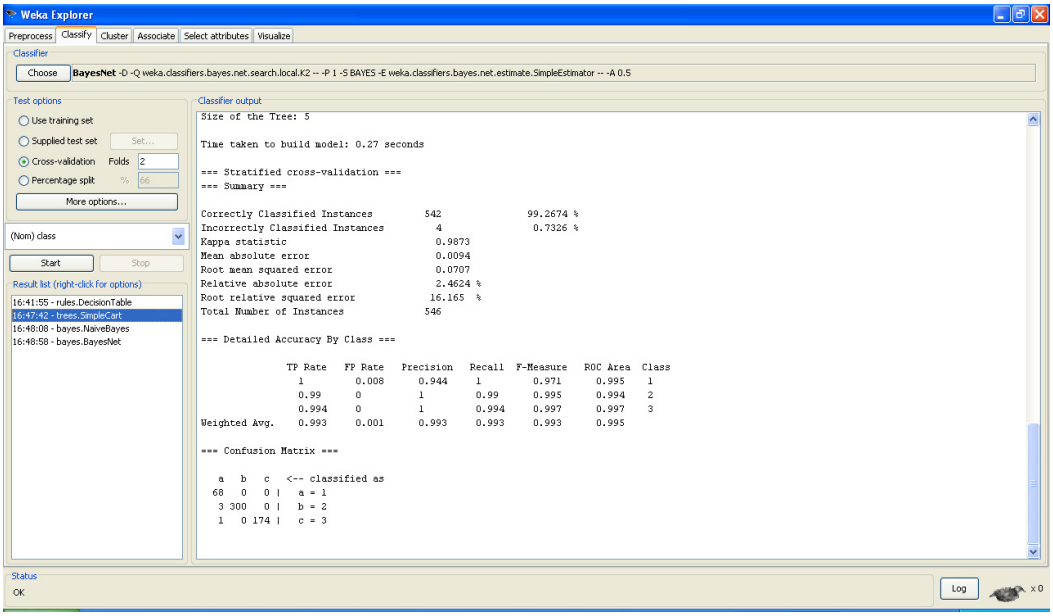

Fig 4 Out put of Cart classification using GRW feature selection method.

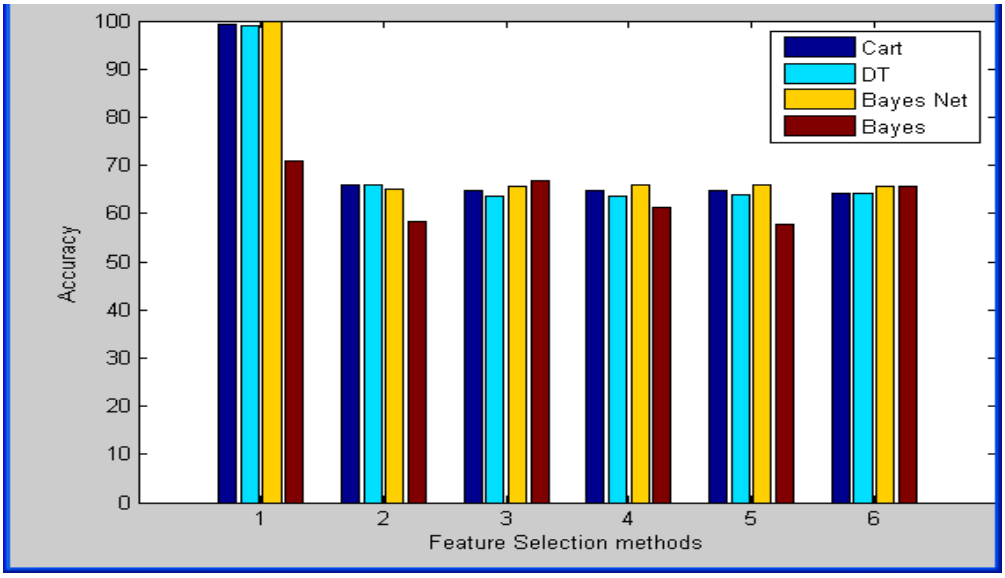

Fig 5 Accuracy of GRW, TFIDF, CFS, Gain ratio, Chi square and filter subset feature selection with various classifiers.

7. CONCLUSIONS

The pubmed repositories are growing at the rate of 500000 articles per year. Classification of Medline documents becomes very complex due to high dimensionality of feature space. In this study we discussed how dimensionality is reduced. Our experiments shows that GRW Feature selection schema used at pre-processing stage improves the performance of Medline abstract Classification. Our algorithm shows that GRW works well in high dimension and unevenly distributed document classification. Only Bayes learning shows less accuracy , but other three learners show high accuracy rate.

REFERENCES

1. Fabrizio Sebastiani, Macine learning in Automated text categorization ,ACM Computing Surveys, VOL34,No 1(2002), pp 1-47
2. J Novovicova et al , Feature selection using Improved Mutual Information for Text classification' ,SSPR & SPR(2004),pp 1010-1017
3. K.Perpinani Why IDF? ,In NAACL 01,Second meeting of the North American Chapter of the Association of Computational Linguistics on Language Technologies (2001), pp 1-8
4. Lecture 2,More Similarity searching Multidimensional scaling 36-350,Data mining ,2009.
5. L.Song A. Smola et al ,Supervsied Feature Selection via dependence estimation , In International conference on Machine Learning 2007
6. Ng et al ,Examining the role of Linguistic Knowledge sources in the automatic identification and classification Reviews, In proceedings of COLING /ACL,2006.
7. Robertson et al, Understanding IDF on theoretical arguments for IDF ,Journal of Documentation ,5:503-520,2004
8. Ronen Feldman, James Sange, The Text mining Handbook, Cambridge University Press(2007).
9. S.Sagar Imambi, T.Sudha - A Unified frame work for searching Digital libraries Using Document Clustering –International Journal of Computational Mathematical ideas Vol 2-No1-(2010) ,pp 28-32
10. Ranjit Abraham et al, Medical Data mining with a new algorithm for Feature selection and Navie Bayesian Classification IEEE 10th International Conference on Information Technology, 2007.
11. S.Sagar Imambi, T.Sudha-Clinical Decision Support System for Heart Patients-International Journal of Computer Science, System Engineering and Information Technology, Vol 2-No2. (2009), pp 165-169
12. Shoushan Li et al , 'A frame work of feature Selection Methods for Text categorization' ,Proceedings of 47th Annual meeting of ACL & 4th ICCNLP of AFNLP (2009), pp 692-700.
13. S.Sagar Imambi, T.Sudha- Classification of Medline documents using Global Relevant Weighing Schema', International Journal of computer Applications 16(3) ,February 2011, pp 45–48
14. Sima C and Dougherty E 'What should be expected from Feature selection in small sample settings ,Bio Informatics 22 (2006), pp 2430-2436

15. S.Sagar Imambi, T.Sudha -.Building Classification System to Predict Risk factors of Diabetic Retinopathy Using Text mining - International Journal on Computer Science and Engineering Vol. 02, No. 07 (2010) ,pp 2309-2312

16. Uğuz H.,A hybrid system based on information gain and principal component analysis for the classification of transactional Doppler signals, Department of Computer Engineering, Selçuk University, Konya, Turkey., 2011 .

17. Yang.y & Pedersen J.O, A comparative study on Feature Selection in Text categorization , 14th Proceedings

18. K. Kira and L.A. Rendell. A practical approach to feature selection. In Sleeman and P. Edwards,editors, Proceedings of the Ninth International Conference on Machine Learning (ICML-92),pp249-256. 1992.
19. H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. In J.F. Vassilopoulos,editor, Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence, November 5-8, 1995, pages 388{391, Herndon, Virginia, 1995. IEEE Computer Society.
20. C. Gini. Variabilite e mutabilita. Memorie di metodologia statistica, 1912.