Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 47 (2015) 76 – 83

Procedia Computer Science

Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis

T. Santhanam a, M.S Padmavathi b

[a] *Associate Professor and Head, PG and Research Department of Computer Science, D. G.Vaishnav College, Chennai - 600106, India.*

[b] *Ph.D Scholor, PG and Research Department of Computer Science, D. G. Vaishnav College,Chennai - 600106, India*

Abstract

Vast amount of data available in health care industry is difficult to handle, hence mining is necessary to find the necessary pattern and relationship among the features available. Medical data mining is one major research area where evolutionary algorithms and clustering algorithms play a vital role. In this research work, K-Means is used for removing the noisy data and genetic algorithms for finding the optimal set of features with Support Vector Machine (SVM) as classifier for classification. The experimental result proves that, the proposed model has attained an average accuracy of 98.79 % for reduced dataset of Pima Indians Diabetes from UCI repository. It also shows that the proposed method has attained better results compared to modified K-Means clustering based data preparation method with SVM classifier (96.71 %) as described in the literature

Peer-review under responsibility of organizing committee of the Graph Algorithms, High Performance Implementations and Applications (ICGHIA2014)

Keywords: Diabetes Diagnosis; Feature Selection; Genetic Algorithms; K-Means; Support Vector Machine.

1. Introduction

Extracting knowledge and patterns for the diagnosis and treatment of disease from the medical database becomes more important to promote the development of telemedicine and community medicine. The methods and applications of medical data include artificial neural network, genetic algorithms, fuzzy

Peer-review under responsibility of organizing committee of the Graph Algorithms, High Performance Implementations and Applications (ICGHIA2014)
doi:10.1016/j.procs.2015.03.185

system, rough set, and support vector machine [1]. One important application of medical mining is genetic algorithms. Dimension reduction is the mapping of data to a lower dimensional space such that uninformative variance in the data is removed, such that a subspace in which the data lives is detected [2]. It can be divided into instance selection or reduction and feature selection techniques. Instance reduction is reducing the irrelevant instances from the dataset to increase the classification accuracy. Selecting a subset of relevant features to be used in model construction is called feature Selection.

Genetic Algorithms (GAs) are intrinsically parallel. GAs have multiple offspring, they can explore the solution space in multiple directions at once. If one path has reached its end, then eliminate it and continue work on other paths, giving them a greater chance each run of finding the optimal solution [3]. There is a greater scope for GAs in future for machine learning and optimization techniques. The K-Means clustering partitions the data into groups which contain similar objects. The instances which do not belong to any cluster (or) a cluster with fewer data points (or) forced to fit into a cluster are removed [4]. Clustering when cascaded with classification considerably increases the accuracy of the model. SVMs are set of related supervised learning methods used for classification and regression. It is a powerful machine method developed from statistical learning and has made significant achievement in the field of image classification, bioinformatics, text categorization, hand-written categorization, etc [5]. Several recent studies have reported that the SVM generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms including, statistical algorithms, decision tree based algorithms and instance based learning methods [6].

This research work proposes K-Means clustering based outlier detection followed by GA for feature selection with SVM as classifier to classify the dataset. The rest of the paper is organized as follows: Section 2 shows the literature review. The preliminaries used is given in section 3 followed by data source description in section 4. The proposed model is described in section 5. Section 6 reports the experimental analysis and the final section deals with a conclusion.

2. Literature Review

Ravi et al [7] proposed a method using GAs to find an appropriate feature subset and SVM classifier on different datasets, to improve classification accuracy. Rajdev et al [8] used a method for selecting optimal feature subset based on correlation using GA algorithm, where GA is used as optimal search tool for selecting subset of attributes. Aishwarya et al [9] proposed a medical decision support system based on GA for feature subset selection and Least Square SVM for diagnosis of diabetes. Shruti et al [10] used GA to find the optimal reduced set of attributes genetic search method and naïve bayes classifier to diagnose the presence or absence of heart disease. Sumeet et al [11] presented a decision support system for heart disease classification based on SVM and integer-coded GA for selecting the important features in diagnosis of heart disease. Sarah Behnam Aziz [12] proposed a method for diagnosis of thyroid diseases using GAs and Neural Networks. The GA was used to find the optimum network structure with high classification accuracy Mehdi et al [13] adopted a hybrid procedure using wrapper subset evaluation and genetic search. Information gain is used to find the optimal feature space and it is proved to be the best method with a lower cost. Ahmad et al [14] applied an improved GA for feature selection and multi-layer perceptron network to classify the medical datasets. Asha et al [15] used clustering and SVM for reducing datasets in the diagnosis of tuberculosis. Hemant et al [16] involved K-Means cluster to reduce datasets with different classification algorithms for predicting diabetes. A hybrid model was proposed by Chin-Yuan Fan et al [17] by integrating a case based data clustering method and a fuzzy decision tree to classify the liver disorder and breast cancer datasets. Nihat et al [18] proposed modified K-Means for removing noisy data and SVM for classification of the reduced datasets. Yuan et al [19] has modeled an objective function which is based on the leave-one-out cross-validation, and the SVM parameters are optimized by using GA and PSO (particle swarm optimization). Patil et al [20] proposed a hybrid K-Means followed followed by naïve bayes and SVM classification, in which SVM achieved high accuracy percentage.

3. Preliminaries

3.1. K-Means Clustering

K-Means clustering algorithm was developed in 1976 by MacQueen. It is a unsupervised clustering algorithm generates a specific number of disjoint, flat (non-hierarchical) clusters The procedure follows a

simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori [21, 22]. K-Means algorithms randomly choses k objects, representing the k initial cluster center. The next step is to take each point belonging to a given data set and associate it to the nearest center based on the closeness of the object with cluster center using Euclidean distance. After all the objects are distributed, recalculate new k cluster centers. The process is repeated until there is no change in k cluster centers. This algorithm aims at minimizing an objective function known as squared error function given by the following.

$$J(v) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (||x_i - v_j||)^2 \quad (1)$$

where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.
'c_i' is the number of data points in i^{th} cluster.
'c' is the number of cluster centers.

The steps of K-Means Algorithm is given below:

1. Randomly partition the dataset into k
2. For each data point in the dataset:
   - *Calculate the distance from the data point to each cluster.*
   - *If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster*.
3. Repeat the above step until a complete pass through all the data points' results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
4. The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intra-cluster distances and cohesion.

3.2. Genetic Algorithms

GA is based on Darwin's theory of natural selection and 'survival of the fittest' [23]. It is one important search method used for feature selection. It represent feasible solutions to the particular problem that evolves over successive iterations (generations) through a process of competition and controlled variation. The process of selection is to select the chromosomes in the population which satisfy the associated fitness to form new ones in the competition process. The genetic operators such as crossover and mutation are used to create the new chromosomes. The classical model of GAs consists of three operations:

1. Evaluation of individual fitness.
2. Formation of a gene pool (intermediate population) through selection mechanism.
3. Recombination through crossover and mutation operators.

The selection mechanism produces a new population ***P(t)*** with copies of chromosomes in ***P(t-1)***. The number of copies received for each chromosome depends on its fitness; chromosomes with higher fitness usually have a greater chance of contributing copies to ***P(t)***. Then, the crossover and mutation operators are applied to ***P(t)***. Crossover takes two individuals called parents and produces two new individuals called the offspring by swapping parts of the parents. In its simplest form, the operator works by exchanging substrings after a randomly selected crossover point. The crossover operator is not usually applied to all pairs of chromosomes in the new population. A random choice is made, where the likelihood of crossover being applied depends on probability defined by a crossover rate. Mutation serves to prevent premature loss of population diversity by randomly sampling new points in the search space. Mutation is applied by flipping one or more random bits in the bit string with a probability equal to the mutation rate. Termination may be triggered by reaching a maximum number of generations or by finding an acceptable solution by some criterion.

3.3. Support Vector Machines

SVM is a classier that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. The SVM method [24] provides an optimally separating hyperplane in the sense that the margin between two groups is maximized. "Support Vectors" are defined as subset of data instances used to define the hyperplane. The distance between the hyperplane and the nearest support vector is called as margin [25]. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. There are two types of SVMs, (a) Linear SVM is used to separate the data points using a linear decision boundary and (b) Non-linear SVM separates the data points using a nonlinear decision boundary [7].

Traditional SVM training algorithms require quadratic programming (QP) package. Solving a quadratic programming problem is slow and requires a lot of memory as well as in depth knowledge of numerical analysis. Consider a binary classification problem with a dataset (x_1, y_1), ..., (x_n, y_n), where x_i is an input vector and $y_i \in \{-1, +1\}$ is a binary label corresponding to it. The dual form of representing quadratic programming problem is given below [11]:

$$\underbrace{max}_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j K(x_i, x_j) \alpha_i \alpha_j , \quad (2)$$

Subject to

$$0 \leq \alpha_i \leq C, \quad for\ i = 1,2, \ldots n, \quad \sum_{i=1}^{n} y_i \alpha_i = 0 \quad (3)$$

Sequential Minimal Optimization does a way with the need for quadratic programming. The Sequential Minimal Optimization (SMO) algorithm (Platt, 1999a) avoids working with numerical QP routines by analytically solving a large number of small optimization sub-problems that involves only two Lagrange multipliers at the time. For any two multipliers α_1 and α_2, the constraints are reduced to the following:

$$0 \leq \alpha_1, \alpha_2 \leq C \quad (4)$$

$$y_1 \alpha_1 + y_2 \alpha_2 = K \quad (5)$$

This reduced problem can be solved analytically by finding a minimum of a one-dimensional quadratic function. The value of K is fixed in each iteration, it is the negative of the sum over the rest of terms in the equality constraint. The algorithm proceeds as follows:

1. Find a Lagrange multiplier α_1 that violates the Karush–Kuhn–Tucker (KKT) conditions for the optimization problem.
2. Pick a second multiplier α_2 and optimize the pair (α_1, α_2).
3. Steps 1 and 2 are repeated until convergence.

The quadratic programming problem has been solved, when KKT [26] is satisfied by the Lagrange multipliers. The above algorithm guarantees convergence and heuristic measures are used to choose the pair of multipliers so as to increase the rate of convergence.

4. Data Source

Pima Indian Diabetes [7, 9, 18] contains female patients with at least 21 years old. It is used to diagnose the presence of diabetes in pregnant women. There are 768 records, out of which 268 cases in class 'positive test for diabetes' and 500 cases for "negative test for diabetes" with 376 records contain missing values. It contains 8 numerical attributes as input and one output variable. The attribute information present in the dataset is as follows:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)

6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 – Absence of disease / 1 – Presence of disease)

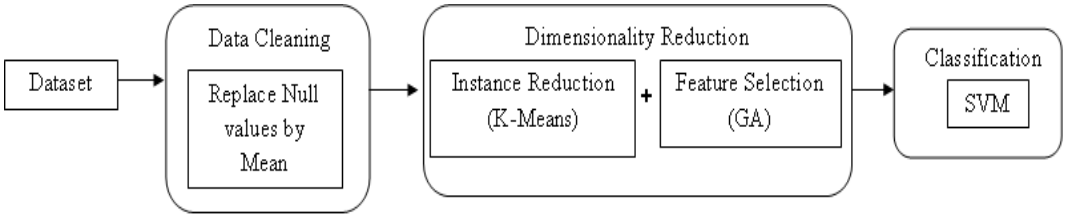

Fig.1 Block Diagram of the Proposed Method

5. Proposed Method

The working principle of proposed system shown in Fig.1, comprises of 3 steps: (1) Data cleaning is done by replacing the missing values with mean (as the dataset lies under a normal distribution curve). (2) The cleaned datasets is clustered using K-Means to remove outliers, inconsistent and noisy data and the reduced data is used for selecting the optimal features with genetic algorithm. (3) The reduced dataset is classified using SVM classifier to achieve better accuracy compared to existing methods in literature. In order to increase the reliability of classifier performance 10-fold cross-validation method is used

5.1. Evaluation Metrics

To visualize the performance of supervised machine learning algorithms, confusion matrix is used. The four classification performance indices present in confusion matrix is given in Fig. 2. The evaluation metrics used are: Sensitivity, Specificity, Positively Predicted and Negatively Predicted value. Sensitivity and Specificity is the proportion of actual positives and negatives which are correctly identified as such. The Positive and Negative predictive values are the proportions of positive and negative results, which are predicted.

$$Sensitivity\ (\%) = \frac{TP}{TP+FN} \times 100 \quad (6)$$

$$Secificity\ (\%) = \frac{TN}{TN+FP} \times 100 \quad (7)$$

$$Positively\ Predicted\ Value\ (\%) = \frac{TP}{TP+FP} \times 100 \quad (8)$$

$$Negatively\ Predicted\ Value\ (\%) = \frac{TN}{TN+FN} \times 100 \quad (9)$$

		Predicted	
		Positive	Negative
rue	Positive	True Positives (TP)	False Negatives (FN)

Fig 2.Confusion Matrix

6. Experimental Results

After replacing the missing values by mean, outliers, noisy and inconsistent samples / cases were removed using simple K-Means clustering algorithm. Then, GA was used as a feature selection tool and its

output was fed to SVM using 10-fold cross validation technique for classification. During each run, GA selected different features from the original set of features and the classification accuracy was recorded. To achieve consistent result, the experiment was repeated 50 times and the outcomes were listed in table 1. The average classification accuracy was 98.79 %.

Among the 50 test runs done, one trail / run which is closer to the average classification accuracy value is chosen to determine the various performance metrics. The evaluation metrics obtained by SVM for classifying the reduced dataset is shown in table 2 and confusion matrix in table 3. From table 4, it is proven that the classification accuracy percentage of the proposed work is better than the existing work.

Table 1. Experimental Outcome for 50 runs.

Test Run	No. of Attributes Selected	SVM Accuracy (%)	Test Run	No. of Attributes Selected	SVM Accuracy (%)
1	5	98.63	26	6	98.43
2	3	98.43	27	5	98.82
3	4	99.21	28	4	98.43
4	5	99.21	29	5	99.02
5	4	99.21	30	4	98.43
6	5	99.02	31	5	98.63
7	6	99.02	32	4	98.43
8	3	98.43	33	4	99.21
9	5	98.43	34	4	98.43
10	5	99.21	35	4	98.82
11	4	98.82	36	4	99.21
12	5	98.82	37	5	99.02
13	6	98.43	38	5	98.63
14	4	99.21	39	5	99.02
15	3	98.43	40	4	99.21
16	4	98.43	41	4	98.82
17	4	98.43	42	4	98.82
18	4	98.63	43	3	98.43
19	4	98.82	44	5	98.82
20	5	99.21	45	5	99.02
21	3	98.43	46	5	99.02
22	4	99.21	47	4	98.82
23	5	98.82	48	5	98.63
24	3	98.43	49	4	99.21
25	5	98.82	50	3	98.43

Table 2. Evaluation Metrics obtained from SVM classifier

Performance Measures	Reduced Dataset
No. of Attributes Used	5
Sensitivity (%)	96.40
Specificity (%)	99.73
Positively Predicted value (%)	99.25
Negatively Predicted value (%)	98.67
SVM Accuracy	**98.82**

Table 3. Confusion Matrix obtained from SVM

	Absent	Present
Absent	134	5
Present	1	371

Table 4. Comparison of the proposed work with the existing works in terms of Accuracy.

Author (Year)	Method	Accuracy (%)
Isa and Mamat (2011) [27]	Clustered-HMLP	80.59
Aibinu et al (2011) [28]	ARI+NN	81.28
Chikh M.A et al (2012) [29]	AIRS2	82.69
	MAIRS2	89.10
Ozcift A. (2012) [30]	RBF+eACC, LADTree+eACC	76.30
Ahmad F. et al (2013) [14]	Improved GA	80.4
Anuja Kumari et al, 2013 [31]	SVM	78
Ravi et al (2014) [7]	GA + SVM	77.3
Nihat Yilmaz et al (2014) [18]	K-Means + SVM	93.65
	Modified K-Means + SVM	96.71
Proposed Method	**K-Means+GA+SVM**	**98.82**

The Research Findings are:

1. The minimum no. of attributes selected using GA is 3 and maximum is 6.
2. The minimum and maximum classification accuracy using SVM is 98.43% and 99.21% and the average accuracy is 98.79 %
3. Out of 768 instances, K-Means selected 511 samples as correctly classified and 257 samples were detected as outliers. The outlier detection percentage is 33.46.
4. Pregnancies, PGConcentration and Age are considered as critical attributes for Pima diabetes dataset.

Conclusion

Medical mining is one major application area where accuracy is important. This study has implemented K-Means and GA for dimensionality reduction and SVM to classify the diabetes dataset. The SVM classification accuracy of the proposed method has an increase of 2.08 % over the Modified K-Means and SVM reported in the literature. The future work will concentrate on use of standard deviation to replace missing values, box plot for outlier detection, algorithms like SVM / PCA for feature selection and to experiment classifiers from statistical, neural, fuzzy and tree families to enhance the resulting accuracy.

References

1. Zhu L1, Wu B, Cao C. Introduction to medical data mining. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi; Sep 2003;20(3); 559-62.
2. Christopher J. C. Burges. Dimension Reduction: A Guided Tour. *Foundations and Trends in Machine Learning*; 2010; 2(4).
3. Gunjan Verma, Vineet Verma. Role and Applications of Genetic Algorithms in Data Mining. *International Journal of Computer Applications (0975–888)*; June 2012;48 (17).
4. Santhanam T, Padmavathi M.S. Comparison of K-Means Clustering and Statistical Outliers in Reducing Medical Datasets. *IEEE:International Conference on Science, Engineering and Management Research (ICSEMR;* Nov 2014; (not yet published).

5. Durgesh K, Srivastava Lekha Bhambhu. Data Classification Using Support Vector Machine. *Journal of Theoretical and Applied Information Technology*; 2009; 12(1).
6. Tony Van Gestel, Johan A. K. Suykens, Bart Baesens, Stijin Viane, Janv Anthienen, Guido Dedene, Bart De Moor, Joos Vandewalle. Benchmarking Least Squares Support Vector Machine Classifiers. Kluwer Academic Publishers; *Machine Learning* ;2004; 54; 5–32.
7. RaviKumar G, Ramachandra G A, Nagamani K, An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets. *International Journal of Advanced Research in Computer Science and Software Engineering;* Feb 2014; 4(2); .272-277.
8. Rajdev Tiwari, Manu Pratap Singh. Correlation-based Attribute Selection using Genetic Algorithm. *International Journal of Computer Applications* (0975–8887); 2010; 4(8).
9. Aishwarya S, Anto S. A Medical Decision Support System based on Genetic Algorithm and Least Square Support Vector Machine for Diabetes Disease Diagnosis. *International Journal of Engineering Sciences & Research Technology*; April 2014; 3(4).
10. Shruti Ratnakar, Rajeshwari K, Rose Jacob. Prediction Of Heart Disease Using Genetic Algorithm For Selection Of Optimal Reduced Set Of Attributes. *International Journal of Advanced Computational Engineering and Networking;April 2013; 1(2);*2320-2106.
11. Sumit Bhatia, Praveen Prakash, Pillai G N. SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features. USA, San Francisco: *Proceedings of the World Congress on Engineering and Computer Science ;* 2008.
12. Sarah Behnam Aziz. Thyroid Disease Diagnosis using Genetic Algorithm and Neural Network. Journal of Qadisiyah Computer Science and Mathematics; 2011; 3 (2); 1-13.
13. Mehdi Naseriparsa, Amir Masoud Bidgoli, Touraj Varaee. A Hybrid Feature Selection Method to Improve Performance of a Group of Classification Algorithms. *International Journal of Computer Applications;* May 2013; *69(17).*
14. .Ahmad F, Isa NA, Hussain Z, Osman MK. Intelligent medical disease diagnosis using improved hybrid genetic algorithm--multilayer perceptron network. *J Med Syst;* April 2013; 37(2).
15. Asha T, Natarajan S, Balasubramanya Murthy K N. A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification. *CORR*; 2011.
16. Hemant P, Pushpavathi T. A novel approach to predict diabetes by Cascading Clustering and Classification. *Computing Communication & Networking Technologies (ICCCNT)*: Third International Conference; July 2012; pp.1-7.
17. Chin-Yuan Fan, Pei-Chann Chang, Jyun-Jie Lin, J.C. Hsieh. A Hybrid model combining Case-based reasoning and Fuzzy Decision Tree for Medical Data Classification. *Applied Soft Computing; 2011; 11(1);* pp.632–644.
18. Nihat Yilmaz, Onur Inan, Mustafa Serter Uzer. A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases. Springer: *Transaction Processing Systems:J Med Syst*; April 2014; 38:48.
19. Yuan R, Guangchen B. Determination of Optimal SVM Parameters by Using Genetic Algorithm/Particle Swarm Optimization. *Journal of Computers*; 2010; No.5;1160-1169.
20. Patil B M, Joshi R C, Toshniwal D. Impact of K-Means on the performance of classifiers for labeled data. *Comm. Com. Inf. Sc*.94;2010; 423-434.
21. Susana A, Leiva-Valdebenito, Francisco J,Torres-Aviles. A Review of the Most Common Partition Algorithms in Cluster Analysis: A Comparative Study. *Colombian Journal of Statistics;* ISSN: 0120-1751; Dec 2010; 33(2); 321-339.
22. Velmurugan T. Efficiency of K-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points . *Int.J.Computer Technology & Applications*; 2012; 3(5); 1758-1764.
23. Goldberg D E. Genetic Algorithms in Search, Optimization and Machine Learning. *Reading,* MA: Addison-Wesley; 1989.
24. Schoslkopf B, Smola A. Learning with Kernels, Support Vector Machines. London: MIT Press; 2002.
25. Vapnik V N.The Natural of Statistical Learning theory. USA, Newyork: Springer–*Verleg*; 1995.
26. John W Chinneck. Practical Optimization: a Gentle Introduction. 2014; http: // www.sce.carleton.ca /faculty /chinneck / po/ Chapter20.pdf.
27. Nor Ashidi Mat Isa, Wan Mohd Fahmi Wan Mamat1. Clustered-Hybrid Multilayer Perceptron network for pattern recognition application. Elsevier: *Applied Soft Computing*; January 2011; 11(1); 1457–1466.
28. Aibinu A M, Salami M J E, Shafie A A. A novel signal diagnosis technique using pseudo complex-valued autoregressive technique. Elsevier: *Expert Systems with Applications;* 2011; 38(8);9063–9069.
29. Chikh M A, Saidi M, Settouti N.Diagnosis of diabetes diseases using an Artificial Immune Recognition System2 (AIRS2) with fuzzy K-nearest neighbor. Springer:*J Med Syst*; Oct 2012;36(5):2721-29.
30. Ozcift A. SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease. Springer:*J Med Syst;* Aug 2012; 6(4):2141-47.
31. Anuja Kumari V, Chitra R. Classification Of Diabetes Disease Using Support Vector Machine. *International Journal of Engineering Research and Applications*; April 2013; 3(2);1797-1801.