Workshop Franco-Brasileiro sobre Mineração de Dados

WFB2009

Workshop Franco-Brasileiro sobre Mineração de Dados

Workshop Franco-Brésilien sur la fouille de données

5-7 mai 2009

Université Fédérale du Pernambouc, RECIFE, Brésil

Centro de Informática - UFPE
Tel +55 81 2126.8430 - Cidade Universitária - 50740-540 - Recife

WFB2009

Workshop Franco-Brasileiro sobre Mineração de Dados

Workshop Franco-Brésilien sur la fouille de données

Université Fédérale du Pernambouc, RECIFE, Brésil

Centro de Informática - UFPE
Cidade Universitária
50740-540 - Recife

Introduction

No quadro do Ano da França no Brasil*, o Centro de Informática - CIn/UFPE e o Conservatoire Nationale des Arts et Métiers (CNAM, França), com o apoio da Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE) e do Institut National de Recherche en Informatique et en Automatique (INRIA, França), organizam o Workshop Franco-Brasileiro sobre Mineração de Dados au Centro de Informática da UFPE.

Este colóquio é consagrado a todas as as problemáticas, teorias, métodos e aplicações da mineração de dados, aprendizagem, extração e a gestão de conhecimentos. O objetivo é unir pesquisadores dessas áreas para apresentar trabalhos de qualidade, realizar trocas e fertilizar novas idéias. Esse domínio de investigação federa trabalhos das áreas de estatística e computação e tem como objetivos a análise e descoberta de estruturas nas grandes bases de dados e fluxos de dados. Eles dizem respeito aos pesquisadores, industriais e empresas interessados em valorizar o conhecimento escondido nas bases de dados ou extrair informações pertinentes nos fluxos de dados.

***França.Br 2009 Ano da França no Brasil** (21 de abril a 15 de novembro 2009) é organizado :

- Na França : pelo Comissariado Geral Francês, pelo Ministério das Relações Exteriores e Européias, pelo Ministério da Cultura e da Comunicação e pelo Culturesfrance.

- No Brasil : pelo Comissariado Geral Brasileiro, pelo Ministério da Cultura e pelo Ministério das Relações Exteriores.

Dans le cadre de l'année de la France au Brésil*, le Centro de Informática - CIn/UFPE et le Conservatoire National des Arts et Métiers (CNAM, France), avec le soutien de la Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE) et de l'Institut National de Recherche en Informatique et en Automatique (INRIA, France), organisent un Workshop Franco-Brésilien sur la Fouille des Données au Centro de Informática de l'UFPE

Ce colloque est consacré à toutes les problématiques, théories, méthodes et applications de la fouille de données, de l'apprentissage, de l'extraction et de la gestion de connaissances. Il s'agit de rassembler les chercheurs de ces domaines afin de communiquer des travaux de qualité, d'échanger et de fertiliser des idées nouvelles. Ce domaine de recherche fédère des travaux de statisticiens et d'informaticiens et a pour objet l'analyse et la découverte de structures dans des grandes bases de données et les flux de données. Sont concernés : chercheurs, industriels et entreprises cherchant à valoriser les connaissances enfouies dans leurs bases de données ou extraire de l'information pertinente dans des flux de données.

***França.Br 2009 » L'Année de la France au Brésil** (21 avril - 15 novembre 2009) est organisée :

- En France : par le Commissariat général français, le Ministère des Affaires étrangères et européennes, le Ministère de la Culture et de la Communication et Culturesfrance.

- Au Brésil : par le Commissariat général brésilien, le Ministère de la Culture et le Ministère des Relations Extérieures.

Remerciements

Les organisateurs remercient chaleureusement les institutions suivantes :

- l'ambassade de France au Brésil pour son soutien financier dans le cadre de l'année de la France au Brésil;
- l'INRIA pour son appui logistique;
- l'association EGC pour son appui financier et son parrainage;
- la SFC et la SFdS pour leur parrainage;
- la FACEPE pour son parrainage;
- le Centro de Informatica - CIn et l'UFPE pour leur appui logistique et financier.

Agradecimentos

Os organizadores agradecem calorosamente as instituições seguintes:

- A embaixada da França no Brasil pelo seu apoio financeiro no quadro do ano da França no Brasil;
- O INRIA pelo seu apoio logístico;
- A associação EGC pelo seu apoio financeiro e seu patrocínio;
- A SFC e a SFdS pelo seu patrocínio;
- A FACEPE pelo o seu patrocínio;
- O Centro de Informatica - CIn e a UFPE para o seu apoio logístico e financeiro.

Scientific Programme Committee

Francisco De Carvalho	UFPE, **Président**
Gilbert Saporta	CNAM, **Co-Président**
Gauss Cordeiro	UFRPE, Permanbouc
Edwin Diday	Paris-Dauphine
Flavio Fogliatto	UFRGS, Rio Grande do Sul
Herman Gomes	UFCG, Paraiba
Georges Hébrail	Telecom-ParisTech
Ludovic Lebart	CNRS and Telecom-ParisTech
Yves Lechevallier	INRIA-Paris-Rocquencourt
Teresa Ludermir	UFPE, Permanbouc
Andre Ponce de Leon de Carvalho	USP, São Paulo
Marcilio Souto	UFRN, Rio Grande do Norte
Emmanuel Viennet	Paris-Nord,
Djamel Zighed	Lyon2-Lumiere

Web Site:

http://www.modulad.fr/Workshop_Franco_Bresilien

Webmaster:

Stéphanie Aubin, INRIA-Paris-Rocquencourt

Contents

Tutorials

Invited Speakers

Contributed Papers

An Introduction to Data Stream Querying and Mining

Georges Hébrail

Telecom-ParisTech
46 Rue Barrault 75013 Paris, France
georges.hebrail@telecom-paristech.fr

Abstract Human activity is nowadays massively supported by computerized systems. These systems handle data to achieve their operational goals and it is often of great interest to query and mine such data with a different goal: the supervision of the system. The supervision process is often difficult (or impossible) to run because the amount of data to analyze is too large to be stored in a database before being processed, due in particular to its historical dimension.

This problem has been studied intensively for several years, mainly by researchers from the database field. A new model of data management has been defined to handle *data streams* which are infinite sequences of structured records arriving continuously in real time. This model is supported by newly designed data processing systems called *Data Stream Management Systems*. These systems can connect to one or several stream sources and are able to process *continuous queries* applied both to streams and standard data tables. These queries are qualified as continuous because they stay active for a long time while streaming data are transient. The key feature of these systems is that data produced by streams are not stored permanently but processed *on the fly*. Note that this is the opposite of standard database systems where data are permanent and queries are transient. Such continuous queries are used typically either to produce alarms when some events occur or to build aggregated historical data from raw data produced by input streams.

As data stored in data bases and warehouses are processed by mining algorithms, it is interesting to mine data streams, i.e. to apply data mining algorithms directly to the streams instead of storing them beforehand in a database. This problem has also been studied a lot and new data mining algorithms have been developed to be applicable directly to streams. These new algorithms process data streams *on the fly* but they can also provide results based on a portion of the stream instead of the whole stream already seen. Portions of streams are defined by fixed or sliding windows.

We will provide an introduction to the data stream management and mining field. First, the main applications which motivated these developments will be presented (telecommunications, computer networks, stock market, security,...) and the new concepts related to data streams will be introduced (structure of a stream, timestamps, time windows,...). A second part will present the main concepts and architectures related to Data Stream Management Systems. The third part will present the main results about the adaptation of data mining algorithms to the case of streams.

Introduction to Text Mining

Ludovic Lebart

CNRS, Telecom-ParisTech
46 Rue Barrault 75013 Paris, France
ludovic.lebart@telecom-paristech.fr

Abstract Principal axes techniques and classification methods play a major role in the computerized exploration of textual corpora. They produce visualizations and/or groupings of elements (free responses in marketing and socioeconomic surveys, discourses, scientific abstracts, patents, broadcast news, financial and economic reports, literary texts, etc.); they highlight associations and patterns; they devise decision aids for attributing a text to an author or a period, for choosing a document within a database, for coding information expressed in natural language. They help also to achieve more technical objectives such as lexical disambiguation, parsing, selection of statistical units, description of semantic graphs, speech and optical character recognition. However, the basic concepts of statistical data analysis must be modified in text analysis. Variables, instead of being declared a priori, are derived from the text. Statistical units (or: observations, subjects, individuals, examples) can be documents (described by their titles or abstracts) in documentary databases, respondents (described by their responses to open questions) in surveys, or segments of texts (sentences, context units, paragraphs) in literary applications. Four additional characteristics increase the complexity of the basic data tables: These tables are large (thousands of documents, thousands of words), often sparse (a document may contain a relatively small number of words) and are provided with a huge amount of available meta-data (rules of grammar, semantic networks). Finally, textual data deal with sequences of occurrences (or: strings) of items, whose order could be of importance, another non standard feature in the multidimensional data analysis. We will focus our presentation on the assessments of visualizations, and the use of meta-data. The examples of application concern open-ended questions in an international survey.

References

[1] Lebart, L., Salem, A., Berry, E.: Exploring Textual Data. Kluwer Academic Publisher, Dordrecht (1998).

Data Mining in Bioinformatics

André C.P.L.F. de Carvalho

Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação e Estatistica
Caixa Postal 668, 13560-970 - São Carlos, SP, Brasil
andre@icmc.usp.br

Abstract Molecular Biology laboratories have gathered a very large amount of data in sequence and functional genome projects. It is frequently not possible to analyze these data manually. Sophisticated computing techniques are necessary to extract new, meaningful and useful information from these data. Data Mining techniques have been successfully applied in such analysis. Examples of these applications are analysis of gene expression data, recognition of genes in DNA sequences and protein structure prediction. This tutorial will present the main issues on the use of Data Mining techniques in Bioinformatics. The tutorial will start with the introduction of the key aspects of Data Mining, with special emphasis on Machine Learning. Next, the necessary issues of molecular biology for the understanding of the Data Mining applications in Bioinformatics will be described. Usually, biological data needs to be pre-processed before they can be used in a Data Mining process. The main techniques for data pre-processing will be presented. Later, a few applications of data Mining techniques, mainly classification and clustering, to bioinformatics problems will be presented.

Transformed Generalized Linear Models

Gauss M. Cordeiro

Departamento de Estatística e Informática, UFRPE
Rua Dom Manoel de Medeiros, s/n - Dois Irmãos
CEP: 52171-900, Recife-PE, Brasil
gauss@deinfo.ufrpe.br

Abstract The estimation of data transformation is very useful to yield response variables satisfying closely a normal linear model. Generalized linear models enable the fitting of models to a wide range of data types. These models are based on exponential dispersion models.We propose a new class of transformed generalized linear models to extend the Box and Cox models and the generalized linear models. We use the generalized linear model framework to fit these models and discuss maximum likelihood estimation and inference. We give a simple formula to estimate the parameter that index the transformation of the response variable for a subclass of models. We also give a simple formula to estimate the $r - th$ moment of the original dependent variable. We explore the possibility of using these models to time series data to extend the generalized auto regressive moving average models discussed by [1]. The usefulness of these models is illustrated in a simulation study and in applications to three real data sets.

References

[1] Benjamin et al. Generalized auto regressive moving average models. J.Amer.Statist. Assoc., (1998) 214–223.

Some Partitioning Clustering Models for Interval-valued Data

Francisco de A.T. de Carvalho

Centro de Informatica - CIn/UFPE
Av. Prof. Luiz Freire, s/n - Cidade Universitária
CEP 50740-540, Recife-PE, Brazil
fatc@cin.ufpe.br

Abstract Cluster analysis have been widely used in numerous fields including pattern recognition, data mining and image processing. Their aim is to organize a set of items into clusters such that items within a given cluster have a high degree of similarity, whereas items belonging to different clusters have a high degree of dissimilarity.

In particular, partitioning clustering models aims to organize a set of items into a pre-defined number of clusters. Our reference clustering model is the partitioning dynamic cluster algorithms [2]. They are iterative two steps relocation clustering algorithms involving at each iteration the construction of the clusters and the identification of a suitable representative or prototype (means, factorial axes, probability laws, etc.) of each cluster by locally optimizing an adequacy criterion between the clusters and their corresponding prototypes.

Often, objects to be clustered are represented as a vector of quantitative features. However, the recording of interval data has become a common practice in real world applications and nowadays this kind of data is often used to describe objects. Symbolic Data Analysis (SDA)is an area related to multivariate analysis, data mining and pattern recognition, which has provided suitable data analysis methods for managing objects described as a vector of intervals [1].

In this presentation, we review partitioning clustering models and algorithms for interval-valued data having as reference the dynamic clustering algorithm. For each clustering model, it is given the clustering criterion, the best ptototype of each cluster, the best distance associated to each cluster (if any) as well as the best partition in a fixed number of clusters. Moreover, various tools for the partition and cluster interpretation of interval-valued data furnished by these algorithms are also presented. Finally, in order to show the usefulness of these algorithms and the merit of the partition and cluster interpretation tools, experiments with real interval-valued data sets are given.

References

[1] Bock, H. H. and Diday, E. (editors) Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data. Springer, Berlin Heidelberg, (2000).

[2] Diday, E. and Simon, J-C., Clustering Analysis. In: K. S. Fu Eds., Digital Pattern Recognition. Springer, Heidelberg, 47–94, (1976).

New Advances in Symbolic Data Analysis and Spatial Classification

Edwin Diday

Ceremade, Université Paris-Dauphine
Place du Maréchal de Lattre de Tassigny
75775 Paris Cedex 16, France
diday@ceremade.dauphine.fr

Abstract The usual Data mining model is based on two parts: the first concerns the units (called here *individuals*), the second, contains their description by several standard variables including a class variable. The Symbolic Data Analysis model needs two more parts: the first concerns units called "concepts" and the second concerns their *description*. The concepts are characterized by a set of properties called "intent" and by an "extent" defined by the set of individuals which satisfy these properties. These concepts are described by *symbolic data* which are standard categorical or numerical data and moreover interval, histograms, sequences of values, etc. These new kind of data allows keeping the internal variation of the extent of each concept. Then, new knowledge can be extracted from this model by new tools of Data Mining extended to concepts considered as new units. Among these tools, Spatial Classification allows a graphical visualisation of the given concepts on a grid and at different level of generalisation organised by a spatial hierarchy or pyramid (allowing overlapping clusters). The SYR software has been developed by SYROKKO company after the academic SODAS software developed by two EUROPEAN projects until 2003, The first aim of SYR is to extract, from a data file (.txt, .csv, ACCESS database) of several millions of units a reduced number of units which are *concepts* summarizing the initial data. Then SYR can create handle (select, cut, move rows or columns.) and visualise a symbolic data file thanks to user-friendly graphical output. Finally SYR produces new knowledge by Symbolic Data Analysis tools.

References

[1] Billard L. and Diday E., Symbolic Data Analysis: conceptual statistics and data Mining. Wiley. ISBN 0-470-09016-2. 351 pages (2006).

[2] Diday E. and Noirhomme M., Symbolic Data Analysis and the SODAS software. Wiley. ISBN 978-0-470-01883-5, 457 pages (2008) .

[3] Diday E., Spatial classification. DAM (Discrete Applied Mathematics) Volume 156, Issue (2008).

SPC-based Strategy for Detecting Frauds in Power Consumption Time Series

Flávio S. Fogliatto

Industrial Engineering and Transportation Department - UFRGS
Praça Argentina, 9, Sala 402,
CEP: 90020-040, Porto Alegre-RS, Brasil
ffogliatto@producao.ufrgs.br

Abstract Non-technical power losses related to fraud and theft are a serious problem in the management of electric power systems, varying in intensity across countries as a function of factors such as effective accountability, political stability, and corruption levels. In the US, where power systems are generally deemed efficient, Nesbit (2000) estimates that nontechnical losses represent from 0.5

The first two modalities may be minimized (i) by investing in metering technology [4] and [3], (ii) through an efficient inspection program [2] and [1], or (iii) by changing system ownership from public to private or by some other market strategy [6]. To accomplish loss minimization through inspection, which is our main concern, one must first focus on the problem of selecting meters to be inspected from a population of consumers such that irregularity detection is maximized. Approaches vary in the literature, although usually dealing with predicting customers' future consumption and analyzing abnormalities in their demand time series. In this paper, we propose an SPC (Statistical Process Control)-based strategy for detecting unusual behavior in customers' demand time series. Although proposing a simplified and easily implementable forecasting model to predict demand, our method is essentially grounded on the analysis of historical demand behavior in search of potential fraudulent customers. For that purpose, we propose the combined use of robust statistics and SPC rules. Our proposal is illustrated in a case study using a large dataset provided by an electricity distributor located in southern Brazil.

References

[1] Ahmad, A.R. and Mohamad, A.M., Intelligent system for detection of abnormalities and probable fraud by metered customers, 19th International Conference on Electricity Distribution, Vienna, 21-24 May 2007.

[2] Cabral, J.E. and Gontijo, E.M., Fraud detection in electrical energy consumers using rough sets, 2004 IEEE International Conference on Systems, Man and Cybernetics, 10-13 Oct. 2004. On CD-ROM.

[3] Ghajar, R. and Khalife, J., Cost/benefit analysis of an AMR system to reduce electricity theft and maximize revenues for Electricite du Liban. Applied Energy, V. 76, 25–37, (2003).

[4] Ghajar, R. and Khalife, J. and Richani, B., Design and cost analysis of an automatic meter reading system for Electricite du Liban. Utilities Policy, V. 9, 193–205, (2000).

Some Clustering Methods on Dissimilarity or Similarity Matrices: Uncovering Clusters in WEB Content, Structure and Usage

Yves Lechevallier

INRIA-Paris-Rocquencourt
78153 Le Chesnay Cedex, France
Yves.Lechevallier@inria.fr

Abstract Clustering is one of the most popular techniques in knowledge acquisition and it is applied in various fields including data mining and statistical data analysis. Clustering involves organizing a set of individuals into clusters in such a way that individuals within a given cluster have a high degree of similarity, while individuals belonging to different clusters have a high degree of dissimilarity.

The definition of *homogeneous cluster* depends on a particular algorithm: this is indeed a structure, which, in the absence of prior knowledge about the multidimensional shape of the data, may be a reasonable starting point towards the discovery of richer and more complex structures.

We propose an clustering method for partitioning a set of objects where the relation between two objects is described by a dissimilarity or similarity measures. The clustering criterion, based on the sum of weighted dissimilarities between the objects belonging to the same class, measures the homogeneity of the cluster. The mathematical properties of these weighted distances and to implement the corresponding algorithms which optimize the clustering criterion and an empirical framework to their evaluation will be studied The advantage of this approach is that the clustering algorithm recognizes different shapes and sizes of clusters.

Clustering is a valuable technique for analyzing the Web. We propose to study clustering approaches in Content and Structure Document Mining and Usage mining. The analysis of a web site based on its usage data is an important task as it provides insight into the organization of the site and its adequacy regarding user needs. We thus defined an approach for discovering the profiles of visitor groups.

An Optimization Methodology for Neural Network Weights and Architectures

Teresa B. Ludermir

Centro de Informatica - CIn/UFPE
Av. Prof. Luiz Freire, s/n - Cidade Universitária
CEP 50740-540, Recife-PE, Brazil
tbl@cin.ufpe.br

Abstract An Optimization Methodology for Neural Network Weights and Architectures This talk introduces a methodology for neural network global optimization. The aim is the simultaneous optimization of multilayer perceptron (MLP) network weights and architectures, in order to generate topologies with few connections and high classification performance for any data sets. The approach combines the advantages of simulated annealing, tabu search and the backpropagation training algorithm in order to generate an automatic process for producing networks with high classification performance and low complexity. Experimental results obtained with four classification problems and one prediction problem has shown to be better than those obtained by the most commonly used optimization techniques. Considering the data sets used in the work presented in this talk, the methodology was able to generate automatically MLP topologies with many fewer connections than the maximum number allowed. The results also generate interesting conclusions about the importance of each input feature in the classification and prediction task. The proposed methodology was originally not designed to deal with different number of hidden layers but it does work with different numbers of hidden layers. Some experiments were made with more than one hidden layer. In any case, a decision needs to be made about the size of the initial topology. So in the experiments made, the initial topologies have only one hidden layer with all possible feedforward connections.

Models for Data or Models for Prediction?

Gilbert Saporta

CEDRIC-CNAM
Conservatoire National des Arts et Métiers
292 rue Saint Martin
75141 Paris cedex 03, France
Gilbert.Saporta@cnam.fr

Abstract The classical view off statistical modelling consists in establishing a parsimonious representation of a random phenomenon, generally based upon the knowledge of an expert of the application field: the aim of a model is to provide a better understanding of data and of the underlying mechanism which have produced it. On the other hand in Data Mining and Statistical Learning predictive models are merely algorithms and the quality of a model is assessed by its performance for predicting new observations. In this communication, we develop some general considerations about both aspects of modelling.

Learning in Social Networks

Emmanuel Viennet

Institut Galilée - Université Paris 13
99, avenue Jean-Baptiste Clément
93430 Villetaneuse, France
emmanuel.viennet@univ-paris13.fr

Abstract The study of Social Networks advanced significantly during last years, with the development of sophisticated techniques for Social Network Analysis and Mining, driven by a strong demand from Web 2.0 applications: social web sites, e-mails and IM systems. The applications includes classification systems (text classification, churn, ...), the detection of users communities and recommendation systems. Social Network Analysis faces difficult problems, like modeling the nature of social interactions, mining structured data (social graphs, text, heterogeneous data), or understanding the dynamic of the social networks. Moreover, the applications typically generate huge datasets, with networks counting several millions of nodes, and the mining algorithm have to deal with the data using limited computing resources. In this communication, we will present several problematics arising in Social Network Analysis, describe some recent advances and give some examples showing how social graph encoding can improve data mining tasks.

Topological approaches in machine learning

Djamel Zighed

ERIC - Université Lumière Lyon 2
5 av. Pierre Mendès-France,
69600 Bron, France
abdelkader.zighed@univ-lyon2.fr

Abstract Plenty of machine learning algorithms have been proposed so far either for supervised or unsupervised learning. Over all, these algorithms stress more on the metrical aspect than on the topological relations between the data set. In this talk we will show why the topological relation is more informative than the metric. We introduce some techniques, based on computational geometry, that show up the topological structure of the data set. Then, we will present some methodological approaches that exploit the topological relationship for the tasks of classification or clustering.

Symbolic Data Analysis Tools for Recommendation Systems

Byron L. D. Bezerra[1], Francisco de A.T. de Carvalho[2]

Departamento de Sistemas e Computação, Universidade de Pernambuco
Rua Benfica, 455 - Madalena - CEP 50720-001 - Recife (PE) - Brazil

Abstract Recommendation Systems have become an important tool to cope with the information overload problem by acquiring data about the user behavior. After tracing the user behavior, through actions or rates, Computational Recommendation Systems use information filtering techniques to recommend items. In order to recommend new items, one of the three approaches has been mainly adopted: Content Based Filtering, Collaborative Filtering or hybrid filtering methods. This paper presents three information filtering methods, each of them based on one of the previous approaches. In our methods, the user profile is designed through Symbolic Data Structures and the user and item correlations are computed through distance functions adapted from the Symbolic Data Analysis domain. The usage of Symbolic Data Analysis tools have improved the performance of Recommendation Systems, specially when there is little information about the user.

Keywords: Symbolic Data Analysis, Recommender Systems, Information Filtering, Information Retrieval.

1 Introduction

Recommendation systems allow e-commerce websites to suggest products to their costumers, providing relevant information to help them in shopping tasks [1], [2], [3]. Additionally, most often this sort of system has increased their importance in entertainment domains [4], [5]. For instance, some interesting features are personalized TV guides in digital televisions and music recommendation in on-line stations.

In order to suggest items, Recommender Systems need user preferences to build suggestions. The process of collecting user preferences may be made implicitly (listening to some music in a CD store or, even better, buying a CD) or explicitly (evaluating some article with a grade in a on-line magazine) [2]. Independently of the acquiring approach (implicit or explicit), as much preferences are collected from user, better suggestions will be provided. But, in fact, a relevant problem remains in this process: the user has not enough time to giving information about him/her. So, it is necessary to learn about user with as little information as possible. This problem is even more difficult to cope with in the first system usage, when there is no information about user. In this case, it is interesting a suitable strategy to acquire user preferences.

The next step is filtering in relevant information in order to present it to the user through his/her profile previously acquired. The proposed solutions for this subject can be classified in two main groups concerning the kind of Information Filtering approach, e. g., Content-based Filtering (which is based on the correlation between the user profile and items content) or Collaborative Filtering (which is based on the users profiles correlation)

[3]. These techniques have inherent limitations, such as impossibility to codify some information in the first approach and latency (or cold-start problem) in the second one. Therefore, several works have exploiting hybrid recommenders to overcome the drawbacks of each [2].

In this paper we describe information filtering techniques which the user profile and item information are modeled by Modal Symbolic (MS) data. This kind of structure was firstly defined in the Symbolic Data Analysis (SDA) field. SDA provides suitable tools for managing aggregated data described by multi-valued variables, where data table entries are sets of categories, intervals, or probability distributions [6].

Based on SDA data structures and tools, we develop three methods. The first one is an evolution of the Content-based approach presented in [7]. The second method may be classified as a Collaborative Filtering approach. Finally, we propose a Hybrid Filtering approach supported by SDA tools.

A deep experimental analysis was carried, taking into account the following issues: (i) the methodology of acquiring preferences, (ii) the algorithm used to learn user preferences, (iii) the amount of information known about the user, (iv) the database size, e. g., the number of users in the community, and (v) the metrics used to evaluate the proposed methods. The experiments were conducted in the movie recommendation domain, where the user profile is formed by way of a list of items that the user either preferred or disliked in the past, along with their respective grades.

The usage of Symbolic Data Analysis tools have improved the performance of Recommendation Systems, specially when there is little information about the user.

References

[1] F. Ricci and H. Werthner, Introduction to the Special Issue: Recommender Systems, *International Journal of Electronic Commerce* **11** (2006) 5–9.

[2] K. Wei, J. Huang and S. Fu, A Survey of E-Commerce Recommender Systems, *International Conference on Service Systems and Service Management* (2007) 1–5.

[3] Z. Huang, D. Zeng and H. Chen, A Comparison of Collaborative-Filtering Recommendation Algorithms for E-commerce, *IEEE Intelligent Systems*, **Vol. 22, N. 5** (IEEE, 2007) 68–78.

[4] Y. Blanco-Fernandez, J.J. Pazos-Arias, M. Lopez-Nores, A. Gil-Solla and M. Ramos-Cabrer, AVATAR: An Improved Solution for Personalized TV based on Semantic Inference, *IEEE Transactions on Consumer Electronics* **52** (2006) 223–231.

[5] J. Salter and N. Antonopoulos, CinemaScreen Recommender Agent: Combining Collaborative and Content-Based Filtering, *IEEE Intelligent Systems* **Vol. 21, N. 1** (2006) 35–41.

[6] H.H. Bock, E. Diday, *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data* (Springer-Verlag, Heidelberg, 2000).

[7] B.L.D. Bezerra, F.A.T. de Carvalho, A Symbolic Approach for Content-based Information Filtering *Information Processing Letters* **92** (2004) 45–52.

Aplicando classificação não-supervisionada para detecção de cola em provas escolares

Elmano R. Cavalcanti[1], José S. Jackson[2]

[1] Universidade Federal de Campina Grande (UFCG), 58.109-970 - Campina Grande - PB - Brazil
[2] Faculdades Integradas de Patos (FIP), 58.700-250 - Patos - PB - Brazil

Abstract Neste artigo, foi desenvolvida uma solução para a detecção de cola em provas escolares utilizando-se de classificação não supervisionada, do modelo de vetor de espaços e do cálculo da similaridade por cosseno. O resultado permitiu detectar cola de diversos tamanhos em um conjunto de 30 provas. A acurácia do modelo de detecção foi de 72,73% e o índice Kappa apresentou um valor substancial: 0,63.

Keywords: classificação, similaridade, mineração de texto

1 Introdução

A cola de alunos em provas escolares é uma prática indesejável porém, amplamente disseminada. Um dos fatores que dificulta a detecção de cola é grande quantidade de dados que são analisados manualmente pelo professor. Embora não exista uma definição concreta de cola, o senso comum de que quando duas provas apresentam um grau de semelhança razoável, supõe-se que tenha ocorrido a cola. Com o advento da informática é de esperar-se que num futuro próximo as correções das provas sejam feitas por *softwares* capazes de detectar diversos tipos e tamanhos de cola, desde as maiores às mais sutis. A seguir, é descrita uma solução para detecção de cola em provas escolares.

2 O Processo de Mineração de Texto

No escopo deste trabalho, os dados foram obtidos diretamente de fontes eletrônicas ou criados manualmente em arquivos de textos comuns. Dessa forma, não foi necessário realizar as atividades iniciais de seleção, limpeza e amostragem dos dados. Ao todo, foram utilizadas 24 provas reais, cada uma contendo quatro questões subjetivas. Adicionalmente, foram criadas seis provas fictícias simulando alunos que colaram de alguma das provas reais.

Foi necessário definir um dicionário para cada questão, ou seja, um conjunto de palavras que faz parte do contexto da questão. Seguindo as etapas de mineração de texto apresentadas em [1], temos as seguintes fases:

1. *CodeMapper*: Nesta etapa, foi realizada a remoção dos acentos das palavras. Todas as etapas a seguir foram feitas utilizando-se a ferramenta RapidMiner [4] com o plugin de mineração texto [3];

2. *Tokenizer*: Separação do texto em palavras (i.e., *tokens*);

cola	grande	razoável	pequena	nenhuma	total
grande	10	2	0	0	12
razoável	1	6	3	0	10
pequena	0	0	4	1	5
nenhuma	0	1	1	4	6
total	11	9	8	5	33

Table 1: Matriz de Confusão

3. *WordFilter*: Em seguida, foi feita uma filtragem da lista de palavras da etapa anterior. Utilizou-se a lista de *stopwords* disponível no projeto Snowball [2]. Foram feitos alguns acréscimos de palavras, de acordo com o domínio das questões da prova;

4. *Stemmer/Reducer*: Antes de iniciar a garimpagem dos dados foi necessário mapear todos os sinônimos ou palavras que possuem o mesmo radical (e.g., processar e processado) para uma única palavra-base. Para esta etapa foi utilizado o algoritmo de *stemming* do Snowball.

5. *Criação do vetor*: O modo usado para criação do vetor foi o *term frequency-inverse document frequency* (TFIDF), que é uma medida estatística utilizada para avaliar a importância que uma palavra tem dentro de um documento. O tamanho do vetor ficou em aproximadamente 500 colunas. Todos os vetores foram normalizados para o tamanho unitário Euclidiano.

Utilizou-se classificação não-supervisionada através do cálculo de similaridade da função cosseno, a qual retorna um valor real $sim(i,j)$ no intervalo [0,1], indicando o nível de semelhança entre duas provas i e j. Foram consideradas quatro categorias de cola: grande ($sim(i,j) > 0,70$), razoável ($0,40 < sim(i,j) < 0,70$), pequena ($0,25 < sim(i,j) < 0,40$), nenhuma ($sim(i,j) < 0,25$). Em seguida, foi possível construir a Matriz de Confusão (Tabela 1), que representou uma acurácia de 72,73% e índice Kappa de 0,63, o que é considerado um valor substancial [5] para o modelo de inferência construído.

References

[1] Bilenko, M., Mooney, R. J.: Adaptive Duplicate Detection Using Learnable String Similarity Measures. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003), Washington DC, August 2003 pp.39–48

[2] http://snowball.tartarus.org/

[3] Wurst, M.: The Word Vector Tool and the RapidMiner Text Plugin. Dortmund, Germany (2007). Disponível em: ¡http://wvtool.sf.net/¿

[4] RapidMiner 4.0 - User Guide, Operator Reference and Developer Tutorial. Dortmund, Germany (2007). Disponível em: ¡http://www.rapidminer.com/¿

[5] Landis, J. R. and Koch, G. G. The measurement of observer agreement for categorical data In Biometrics, vol. 33, 1977, pp. 159–174

Classification of Clinical Time Series with Constrained estimation of Mixtures of HMMs

Ivan G. Costa[1], Alexander Schönhuth[2], Christoph Hafemeister[3] Alexander Schliep[3]

[1] Center of Informatics, Federal University of Pernambuco, Recife, Brazil
[2] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada
[3] Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany

Abstract The use of molecular aspects of diseases for clinical diagnosis, has become increasingly popular. There are several difficulties in analyzing clinical gene expression data; high dimension of feature spaces vs. few examples, noise and missing data. We use constrained estimation of mixtures of hidden Markov models as a methodology to classify Multiple Sclerosis patient response to IFNβ treatment. The approach models the temporal nature of the data, is robust to noise and mislabeled examples. It also allows finding sub-groups of patients. Our method outperforms all previously method, and indicate the existence of biologically interesting sub-groups of patients.

Keywords: contrained mixture estimation, hidden Markov Models, gene expression time courses, clinical diagnosis

1 Introduction

The use of molecular aspects of diseases for clinical diagnosis, the so called personalized medicine, has become increasingly popular. One faces multiple challenges when analyzing clinical gene expression data; most of the well-known theoretical issues such as high dimension of feature spaces vs. few examples, noise and missing data apply. Special care is needed when designing classification procedures that support personalized diagnosis and choice of treatment. We analyse here the classification of interferon-β (IFNβ) treatment response in Multiple Sclerosis (MS) patients. Half of the patients remain unaffected by IFNβ treatment, which is still the standard. For them the treatment should be timely ceased to mitigate the side effects.

2 Method

We investigate the problem of classification of Multiple Sclerosis (MS) patients with respect to their response to Interferon-beta (IFNβ) treatment based on their gene expression profiles alone. IFNβ can still be considered to be the standard treatment in MS [Baranzini 2005]. To classify and further explore clinical differences between the groups of good and bad responders [Baranzini 2005], followed fifty-two patients for two years after initiation of IFNβ therapy. Every three months expression profiles of 70 genes were measured. Patients were divided into good and bad responders based on clinical criteria such as relapse rate and disability status. They demonstrated that the patients' response could be predicted by studying gene expression profiles of the first time point after treatment alone [Baranzini 2005].

We propose constrained estimation of mixtures of hidden Markov models as a methodology to classify patient response to IFNβ treatment. By using HMMs with linear topology, our method takes the temporal nature of the data into account and allows the modelling of patient specific response rate [Schliep 2005]. Moreover, constraint based mixture estimation enables to explore the presence of response sub-groups of patients based on their expression profiles and allows the detection of misllabelled samples.

3 Results

We perform a 5 replications 4 fold application procedure and measure the classification accuracy in the test sets. Our method had a accuracy of 92% [Costa 2009]. It outperformed all prior approaches [Baranzini 2005, Borgwardt 2006, Lin 2008], which had a maximun acuracy of 88%. Additionally, we were able to identify potentially mislabeled samples, which was latter confirmed by the authors of the original paper [Baranzini 2005]. Furthermore, our results subdivide the good responders into two subgroups that exhibited different transcriptional response programs. This is supported by recent findings on MS pathology and therefore may raise interesting clinical follow-up questions.

References

[Baranzini 2005] Baranzini, S. E., Mousavi, P., Rio, J., Caillier, S. J., Stillman, A., Villoslada, P., Wyatt, M. M., Comabella, M., Greller, L. D., Somogyi, R., Montalban, X., and Oksenberg, J. R. (2005). Transcription-based prediction of response to ifnbeta using supervised computational methods. *PLoS Biol*, **3**(1), e2.

[Borgwardt 2006] Borgwardt, K. M., Vishwanathan, S. V. N., and Kriegel, H.-P. (2006). Class prediction from time series gene expression profiles using dynamical systems kernel. *Pacific Symposium on Biocomputing*, **11**, 547–558.

[Costa 2009] Costa, I. G., Schönhuth A., Hafemeister C., Schliep A. (2009) Constrained Mixture Estimation for Analysis and Robust Classification of Clinical Time Series *Bioinformatics*, To appear.

[Lin 2008] Lin, T. H., Kaminski, N., and Bar-Joseph, Z. (2008). Alignment and classification of time series gene expression in clinical studies. *Bioinformatics*, **24**(13), i147–i155.

[Schliep 2005] Schliep, A., Costa, I. G., Steinhoff, C., and Schönhuth, A. (2005). Analyzing gene expression time-courses. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **2**(3), 179–193.

Symmetrical Linear Regression models for Symbolic Interval Data

Marco A. O. Domingues[1], Renata M.C.R. de Souza[1],
Francisco José A. Cysneiros[2]

[1] Centro de Informática
[2] Departamento de Estatística,
Universidade Federal de Pernambuco, P.O.Box 7851, Recife (PE), Brazil

Abstract This paper introduces a symmetrical linear regression model as an approach to fit a linear regression for Interval Valued Data.

Keywords: Symmetrical models, symbolic data.

1 Introduction

Symbolic Data Analysis (SDA) could be broadly defined as an extension of standard data analysis to symbolic data. In terms of Regression Analysis, recent works have been proposed to fit the classic linear regression model (CLRM) to symbolic Interval Valued Data (IVD) [1][2]. Those approaches do not consider any probabilistic hypothesis on the response variable and use least squares method to perform parameter estimates whose results are strongly influenced by the presence outliers.

This work introduces a new prediction method for IVD based on the symmetrical linear regression (SLR) analysis. Its main feature is that the response model is less susceptible to the presence of IVD outliers. The model considers the Student-t distribution as an assumption for the errors in the centre of the symbolic interval variables.

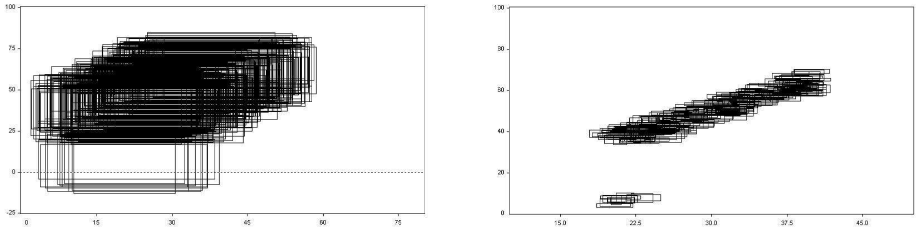
2 Symmetrical linear regression

The SLR model is defined as $Y_i = \mu_i + \epsilon_i, \quad i = 1, \ldots, n$, where $\mu_i = \mathbf{x}_i^t \boldsymbol{\beta}$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ is an unknown parameters vector, $\epsilon_i \sim S(0, \boldsymbol{\phi}, g)$ and $\mathbf{x}_i$ is the vector of explanatory variables. This class of models includes all symmetric continuous distributions, such as normal, Student-t, logistic, among others. The maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ cannot be obtained separately and closed-form expressions for this estimates do not exist. *Scoring Fisher* method can be applied to get $\hat{\boldsymbol{\phi}}$ where the process for $\hat{\boldsymbol{\beta}}$ can be interpreted as a weighted least square. The iterative process for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\phi}}$ takes the form $\boldsymbol{\beta}^{(m+1)} = \{\mathbf{XD}(\mathbf{v}^{(m)})\mathbf{X}\}^{-1}\mathbf{X}^t\mathbf{D}(\mathbf{v}^{(m)})\boldsymbol{y}$ and $\phi^{(m+1)} = \frac{1}{n}\{\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\}^T\mathbf{D}(\mathbf{v})\{\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\}(m = 0, 1, 2, \ldots)$ with $\mathbf{D}(\mathbf{v}) = \mathrm{diag}\{v_1, \ldots, v_n\}$, $\boldsymbol{y} = (y_1, \ldots, y_n)^t$, $\mathbf{X} = (\mathbf{x}_1^t, \ldots, \mathbf{x}_n^t)^t$ and $v_i = -2\mathrm{W}_g(u_i)$, $\mathrm{W}_g(u) = \frac{g'(u)}{g(u)}$, $g'(u) = \frac{dg(u)}{du}$ and $u_i = (y_i - \mu_i)^2/\phi$.

For the Student-t distribution with ν degrees of freedoms, $g(u) = c(1+u/\nu)^{-(\nu+1)/2}, \nu > 0$ and $u > 0$ so that $\mathrm{W}_g(u_i) = -(\nu+1)/2(\nu+u_i)$ and $v_i = (\nu+1)/(\nu+u_i), \forall i$. In this case the current weight $v_i^{(r)}$ is inversely proportional to the distance between the observed value y_i and its current predicted value $\mathbf{x}_i^t\boldsymbol{\beta}^{(r)}$, so that outlying observations tend to have small weights in the estimation process [3].

3 Experiments

To show the usefulness of SLRM, a small subset of the simulated IVD are clustered and changed into outliers by moving the centre of each observation 1. In order to analyze the proposed method we performed Monte Carlo simulations with 100 iterations considering each data set and their simulated outliers.



(a) *Data set* 1.

(b) *Data set* 2.

Figure 1: Interval-valued data sets containing outlier rectangles.

The performance assessment of the SLR model presented is based on the *pooled root mean-square error* ($PRMSE$) applied for a learning IVD set (n=250) and test IVD set (n=125). $PRMSE^1 = \sqrt{\frac{\sum_{i=1}^{250} \omega(i)[(l_Y(i)-\hat{l}_Y(i))^2+(u_Y(i)-\hat{u}_Y(i))^2]}{250}}$, where $\omega(i)$ is the weight of the residual obtained from SLRM and $PRMSE^2 = \sqrt{\frac{\sum_{i=1}^{125}[(l_Y(i)-\hat{l}_Y(i))^2+(u_Y(i)-\hat{u}_Y(i))^2]}{125}}$

Statistical Student's t-test for paired samples at a significance level of 1% is then applied to compare the proposed SLR model to the linear regression model for IVD. The hypotheses are, respectively: $H_0 : (PRMSE^k)^{Symmetrical} = (PRMSE^k)^{Linear}$ and $H_1 : (PRMSE^k)^{symmetrical} < (PRMSE^k)^{Linear}$.

For all test data sets in this evaluation the rejection ratios of H_0 are equal to 100%.

4 Conclusions

A symmetrical linear prediction model for symbolic IVD is introduced in this paper, and experiments with simulated IVD sets containing IVD outliers are performed. The prediction performance is assessed by a PRMSE applied to learning and test data sets, and results provided by the proposed method are compared to the correspondent results provided by least squares method.The results showed that the symmetrical model is superior to centre-range model in terms of prediction qualities.

References

[1] Billard, L., Diday, E., 2006. Symbolic Data Analysis: Conceptual Statistics and Data Mining, Wiley, West Sussex, England 2006.

[2] Lima Neto, E.A., De Carvalho, F.A.T., 2008. Centre and Range method for fitting a linear regression model to symbolic interval data. In CSDA, v.52 n.3, pp. 1500-1515.

[3] Cysneiros, F.J.A., Paula, G.A., 2005. Restricted methods in symmetrical linear regression models. In CSDA, v. 49, pp 689-708.

A Receptive and Inhibitory Field Based Approach for Image Processing

Bruno J. T. Fernandes[12], George D. C. Cavalcanti[1], Tsang I. Ren[1]

[1] Center of Informatics - Federal University of Pernambuco, Brazil
[2] Serra Talhada Unit - Rural Federal University of Pernambuco, Brazil

Abstract This paper presents some models developed to accomplish a pattern recognition task using the concepts of receptive and inhibitory fields. Some discussions about these concepts and its applications are made. Finally, it is presented some future works.

Keywords: Pattern recognition, image processing, neural networks, receptive fields

At the beginning of the 60's, an important region of neurons in the human brain was found. Such region was called receptive field and its actuation was presented in many functionalities of the human brain, it has already been identified in many parts of the human brain, like the auditory, somatosensory and visual systems.

A receptive field is defined by Levine and Shefner [1] as an area in which the presence of an appropriate stimulus might lead to a response of a sensitive neuron.

Many works were presented using the concepts of receptive fields in image processing. A very important one is the work done by Phung and Bouzerdoum [2], where they proposed a new artificial neural network called PyraNet, proposed to accomplish visual pattern recognition tasks. The PyraNet gets as input an image 2-D and outputs its classification. Such neural network is composed in its basis by 2-D layers of neurons, where each neuron is connected to a receptive field in the previous layer. The 2-D layers are responsible by the feature extraction and data reduction of the image. On the top of the PyraNet the neurons are organized in 1-D layers and are responsible by the pattern classification.

In the work done by Phung and Bouzerdoum [2], the PyraNet was applied over a face detection and a gender recognition tasks. The PyraNet obtained good results in such tasks, achieving a high classification tax with short processing time and consuming little memory.

On the other hand, in the work made by Rizzolatti and Camarda [3] was demonstrated that another stimulus, simultaneously to the receptive field stimulus, can also have effect over the sensitive neuron. Such stimulus was called the non-classical receptive field (non-CRF), having an inhibitory effect in most of the times, then we might refer to it as inhibitory field.

The inhibitory field was successfully applied in a contour detection task by Grigorescu [4].

Then, taking in consideration that the PyraNet only consider the excitatory stimulus of the receptive fields, Fernandes et al. [5] proposed the I-PyraNet, a hybridization between the PyraNet and the concepts of inhibitory fields.

The main benefit of the use of the inhibitory concepts in the I-PyraNet is that a same neuron will be able to produce two different outputs according to its spatial position. The presence of a neuron in a inhibitory field will affect only its output, inverting its signal. It is important to note that the PyraNet might be considered a special case of the I-PyraNet where all the sizes of the inhibitory fields are equal to 0.

Also, in the work done by Fernandes et al. [5] it was proposed a supervised image segmentation model based on the concepts of receptive fields, called Segmentation and Classification with Receptive Fields. Then, both, the SCRF model and the I-PyraNet, were applied together with successful in the accomplishment of a forest detection task in satellite images. In this case, the I-PyraNet reached the lowest error rate among all the classifiers tested, including the combination between the SCRF model and the PyraNet.

The I-PyraNet was also applied over a face detection [6] task and obtained, one more time, a better classification rate than the PyraNet. However, in this application the best result was achieved with a Support Vector Machine (SVM). While the I-PyraNet achieved an area of 0.83 In a ROC curve, the SVM achieved an area of 0.9. Although the I-PyraNet have reached a worse error rate than the SVM, the I-PyraNet classification speed was 175 times faster than the SVM, taking only 0.04 milliseconds to classify a pattern. This fact might motivate the use of the I-PyraNet in embedded systems, where the processing time is a big issue in comparison with other kinds of systems. On the other hand, the results of the I-PyraNet can be improved if the classification gets repeated many times without interfering very much in the total classification time in comparison to the SVM.

However, in none of the works cited above was proposed any method to find out the best I-PyraNet configuration, including its receptive and inhibitory fields sizes. Such research field is one of the future works that are going to be studied by us.

References

[1] M. Levine and J. Shefner, *Fundamentals of sensation and perception.* Oxford University Press, 2000.

[2] S. L. Phung and A. Bouzerdoum, “A pyramidal neural network for visual pattern recognition,” *IEEE Transactions on Neural Networks*, vol. 18, no. 2, 2007.

[3] G. Rizzolatti and R. Camarda, “Inhibition of visual responses of single units in the cat visual area of the lateral suprasylvian gyrus (clare-bishop area) by the introduction of a second visual stimulus,” *Brain Res.*, vol. 88, no. 2, pp. 357–361, 1975.

[4] C. Grigorescu, N. Petkov, and M. A. Westenberg, “Contour detection based on nonclassical receptive field inhibition,” *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 729–739, 2003.

[5] B. J. T. Fernandes, G. D. C. Cavalcanti, and T. I. Ren, “Nonclassical receptive field inhibition applied to image segmentation,” *Neural Network World*, vol. 19, pp. 21–36, 2009.

[6] B. J. T. Fernandes and G. D. C. Cavalcanti, “A pyramidal neural network based on nonclassical receptive field inhibition,” *20th IEEE International Conference on Tools with Artificial Intelligence*, pp. 227–230, 2008.

A Comparative Study of Kernel and Classical Methods in Supervised Learning

Marcelo R. P. Ferreira[1], Getúlio José Amorim do Amaral[2]

[1] Departamento de Estatística, CCEN, UFPB
[2] Departamento de Estatística, CCEN, UFPE

Abstract Methods based on kernel density estimation have been used in a wide variety of real-world discrimination problems. This work reviews some classical statistical methods that are frequently used in supervised learning: logistic regression, k-Nearest Neighbour, normal based linear and quadratic classifiers; and a non-parametric one: the kernel classifier. Applications with real data sets are used to compare the classification methods. Our results show that the kernel method outperforms the classical approachs in many situations.

Keywords: Kernel density estimation, Kernel density classification, Classification, Misclassification rate.

1 Introduction

Consider data $\{\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n\}$, where $\underset{\sim}{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, as a realization of a random sample, and let an element of the set $\{f_j(\underset{\sim}{x}), j = 1, \ldots, J\}$ be the density associated with $\underset{\sim}{x}_i$. Let π_j, $j = 1, \ldots, J$, be the classes' prior probabilities, *i.e.* $\pi_j = P(\underset{\sim}{x}_j \in \Pi_j)$ where Π_j denotes the jth class. Then, using Bayes' Theorem, the posterior probability of the observation $\underset{\sim}{x}_i$ being from the jth class, is

$$P(\underset{\sim}{x}_j \in \Pi_j | \underset{\sim}{x}_i = \underset{\sim}{x}) = \frac{\pi_j f_j(\underset{\sim}{x})}{\sum_{j=1}^{J} \pi_j f_j(\underset{\sim}{x})}.$$

According to Bayes' formula, we allocate an observation to the class with highest posterior probability:

$$\underset{\sim}{x} \text{ is allocated to the class } \Pi_j \text{ if } \Pi_j = \underset{j \in \{1, \ldots, J\}}{\arg\max}\ \pi_j f_j(\underset{\sim}{x}).$$

Often the prior probabilities π_j are known, or simply estimated using $\hat{\pi}_j = n_j/n$, $j = 1, \ldots, J$, with $\sum_{j=1}^{J} n_j = n$. Classical parametric approachs make assumptions about the densities f_j. Usually, the data is assumed to have a normal distribution, however, this assumption is very restrictive. With non-parametric discriminant analysis we relax this assumption and thus are able to tackle more complex cases.

The kernel approach for discrimination is to estimate the density f_j of each class Π_j and allocate an observation according to the rule:

$$\underset{\sim}{x} \text{ is allocated to the class } \Pi_j \text{ if } \Pi_j = \underset{j \in \{1, \ldots, J\}}{\arg\max}\ \hat{\pi}_j \hat{f}_j(\underset{\sim}{x}),$$

where $\hat{f}_j(\underset{\sim}{x})$ is the kernel density estimate corresponding to the jth class.

The kernel density estimator of f at the point $\underset{\sim}{x} \in \mathbb{R}^p$ is (see [1, 3] for further details)

$$\hat{f}(\underset{\sim}{x}) = \hat{f}(\underset{\sim}{x}; \boldsymbol{H}) = n^{-1} \sum_{i=1}^{n} K_{\boldsymbol{H}}(\underset{\sim}{x} - \underset{\sim}{x}_i),$$

where the scale factor $\boldsymbol{H}$ is a symmetric positive definite $p \times p$ matrix called the smoothing parameter or bandwidth matrix, and $K_{\boldsymbol{H}} = |\boldsymbol{H}|^{-1} K(\boldsymbol{H}^{-1}\underset{\sim}{x})$, where $K : \mathbb{R}^p \to \mathbb{R}$ is called the kernel; usually K is a symmetric probability density function.

2 Numerical Results

In this section, we will present some numerical results with real and simulated data sets. The real data set (labelled "salmon data") obtained from [2] contains information on growth ring diameters (freshwater and marine water) of 100 salmon fish coming from Alaskan or Canadian water. A random sample ($n = 50$) was used as training sample and the remaining observations were used as test sample. The simulated data set (labelled "synthetic data") is a two-class classification problem. We generate training and testing samples, both with size $n = 100$, from normal mixture densities:

$$\Pi_1 : f_1 \sim \frac{1}{2} N \left(\begin{bmatrix} -\frac{3}{2} \\ -\frac{3}{2} \end{bmatrix} ; \begin{bmatrix} \frac{4}{5} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{4}{5} \end{bmatrix} \right) + \frac{1}{2} N \left(\begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} ; \begin{bmatrix} \frac{4}{5} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{4}{5} \end{bmatrix} \right)$$

$$\Pi_2 : f_2 \sim \frac{1}{2} N \left(\begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \end{bmatrix} ; \begin{bmatrix} \frac{4}{5} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{4}{5} \end{bmatrix} \right) + \frac{1}{2} N \left(\begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix} ; \begin{bmatrix} \frac{4}{5} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{4}{5} \end{bmatrix} \right)$$

The kernel discriminant (KD) was compared with the normal linear (LD) and quadratic discriminants (QD). The misclassification rates are shown in the following table. The results show that the kernel classifier have better performance than the other methods.

Table 1: Misclassification rates on test samples

	Misclassification rate (%)	
Discriminant	salmon data	synthetic data
LD	10	31
QD	10	35
KD	08	19

References

[1] Duong, T.: Bandwidth Selectors for Multivariate Kernel Density Estimation. PhD Thesis, University of Western Australia, School od Mathematics and Statistics. (2004)

[2] Johnson, R. A. & Wichern, D. W.: Applied Multivariate Statistical Analysis. Prentice-Hall, New York. (1998)

[3] Scott, D. W.: Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons, New York. (1992)

Application of Information Extraction in Large Semi-Structured Text

Valmir Macário Filho,[1], Ricardo B. C. Prudêncio[1], Francisco A. T. De Carvalho[1], Leandro R. Torres,[2], Laerte Rodrigues Júnior[2]

[1] Center of Informatics, Federal University of Pernambuco, Av. Prof. Luiz Freire, s/n
50740-540 Recife/PE BRAZIL
[2] Capital Login, R. da Guia, N 99 50.030-210 Recife/PE BRAZIL

Abstract Information extraction systems are used to extract only relevant text information in digital repositories. The current work proposes an automatic system to extract information in semi-structured official journals. The implemented system deployed different features sets and algorithms used in the classification of the fragments. The system was evaluated through experiments on a sample containing 22770 lines of the Pernambuco's Official Journal. The experiments performed revealed, in general, good results in terms of precision.

Keywords: semi-structured document, information extraction, text mining

1 Information Extraction on Official Journal

Official journals are documents that contain publications (e.g., acts, texts of new laws, edicts, decisions) of countries, states, cities and other institutions in the different branches of Executive, Legislative and Judiciary power. The task of finding specific information of interest in official journals is very difficult due to the great number of publications which are daily available. Although this task can be automated, it is possible to point out some difficulties with regard to this purpose: the lack of rigid models to organize the publications in the documents, no clear delimiters between different publications, the presence of abbreviated words, the presence of orthographic errors, among others.

Documents which present the above-cited characteristics are called semi-structured texts [1]. In order to manipulate such documents, an automatic system called Information Extraction (IE) system may be very suitable. IE systems are able to extract specific information of interest from a repository of textual documents. Each input of an IE system is a textual document and the output is a set of text fragments which correspond to data fields required by the user. The extracted fields can be either directly presented to the user or stored in a database for posterior access [2].

The current work develops an IE system to extract information from official journals by using ML algorithms. A publication of an official journal is a semi-structured text divided into five main fields: title, sub-title, notebook, city and process. Some difficulties to extract information from official journals can be mentioned here: (1) fields may present very similar patterns (e.g., the sentence "Edital de Intimação", may appear in the beginning of both fields subtitle and process); (2) absent fields; and (3) presence of abbreviated patterns (e.g., the word "Process" is in many publications abbreviated to "Proc.").

The architecture of the IE system which deployed the text classification approach for IE has three steps:

1. *Fragmentation*: the input text is broken into fragments which are the candidates for filling in the required data fields. In our domain, the fragments correspond to the text lines.

2. *Feature extraction*: a vector of features is created to describe each text fragment and it is used in the classification of the fragment. This task was accomplished by considering a domain vocabulary, regular expressions and text formatting features.

3. *Fragment classification*: a learned classifier associates each input fragment to a class label associated to a data field. In our system, there are ten possible class labels. In this step, we evaluated the use of three classifiers, each one representing a different family of learning algorithms: (1) the PART algorithm for inducing decision rules, (2) the Naive Bayes classifier and (3) the Support Vector Machine (SVM) classifier [3].

In our work, the performed experiments were based on a corpus of publications collected from the Judiciary segment of the Official Journal published by the State of Pernambuco, Brazil.The performance of the IE system was evaluated for 21 different scenarios (i.e., different combinations of feature sets *versus* classifiers). The same above scenarios were also applied to evaluate the usefulness of the Sliding Window (SW) approach. The experiments performed revealed, in general, good results in terms of precision, which ranged from 70.14% to 98.63% depending on the feature set and algorithm used in the classification of the fragments.

The implemented system deployed the text classification approach for IE, which revealed to be adequate for our purpose. We highlight that the application of text classifiers for IE in the domain of Official Journals is an original work. In our experiments, we evaluated different features sets and learning algorithms in the classification of the text fragments. We observed that an improvement in performance can be yielded when sequential information of the fragments is taken into account.

The IE system can be extended to other domains of application. Additionally, as future work, other approaches can be used to construct the feature sets. We also intend to use evaluate sequential learning algorithms, such as Hidden Markov Models and Conditional Random Fields to classify the fragments.

References

[1] Turmo, J., Ageno, A. and Català, N. Adaptive Information Extraction. *ACM Computing Surveys* 38(2):4 2006.

[2] Appelt, D. and Israel, D. Introduction to Information Extraction Technology. *IJCAI-99 Tutorial*, Stockholm, Sweden, 1999.

[3] Duda, R.O. and Hart, P.E. and Stork, D.G. Pattern Classification. *John Wiley & Sons*, 2001.

Complementary log-log and probit: news activation functions in the multilayer perceptron networks

Gecynalda S. da S. Gomes and Teresa B. Ludermir[1]

Centro de Informática, Universidade Federal de Pernambuco, Av. Prof. Luiz Freire - s/n - Cidade Universitária - CEP 50740-540 - Recife (PE) - Brazil

Abstract The types of activation functions most often used in artificial neural networks are logistic and hyperbolic tangent. Activation functions used in ANN have been said to play an important role in the convergence of the algorithms used. This paper uses sigmoid functions in the processing units of neural networks. Such functions are commonly applied in statistical regression models. The nonlinear functions implemented here are the inverse of complementary log-log and probit link functions. A Monte Carlo framework is presented to evaluate the results of prediction power with these nonlinear functions.

Keywords: Activation Function, Multilayer Perceptron Networks, Complementary log-log, Probit, Monte Carlo.

1 Introduction

Artificial neural networks (ANN) may be used as an alternative method to binomial regression models for binary response modelling. The binomial regression model is a special case of an important family of statistical models, namely Generalized Linear Models (GLM) [7, 8]. Briefly outlined, a GLM is described by distinguishing three elements of the model: the random component, the systematic component and the link between the random and systematic components, known as the link function. According to the distribution proposed for the data, the choice of link function may facilitate the interpretation of the model. The majority of current neural network models use the logit activation function, but the hyperbolic tangent and linear activation functions have also been used. ANN have the ability to map nonlinear relationships without a priori information on the process or the system model. More details in [1, 5].

A number of different types of functions have been proposed. Hartman *et al.* (1990) [4] proposed *gaussian bars* as a activation function. *Rational transfer functions* were used by Leung and Haykin (1993) [6] with very good results. Singh and Chandra (2003) [9] proposed a class of sigmoidal functions that were shown to satisfy the requirements of the universal approximation theorem (UAT).

The aim of our work is to implement sigmoid functions commonly used in statistical regression models in the processing units of neural networks and evaluate the prediction performance of neural networks. The functions used are the inverse functions of the following complementary log-log and probit link functions, respectively: $g(\pi) = \ln[-\ln(1-\pi)]$ and $g(\pi) = \Phi^{-1}(\pi)$, in which $g(\cdot)$ denotes the link function and $\Phi(\cdot\,)$ denotes the cumulative probability function for the normal distribution.

We use multilayer perceptron (MLP) networks. The calculations made for the outputs $y_i(t) = \phi_i(\mathrm{w}_i^\top(t)\mathrm{x}(t)), \quad i = 1, \ldots, q$, such that w_i is the weight vector associated with the node i, $\mathrm{x}(t)$ is the attribute vector and q is the number of nodes in the hidden layer. The activation function ϕ is given by one of the following forms: $\phi_i(u_i(t)) = 1 - \{\exp[-\exp(u_i(t))]\}$, represent the complementary log-log and $\phi_i(u_i(t)) = \Phi(u_i(t))$, represent the probit. The derivatives form of the complementary log-log and probit are $\phi_i'(u_i(t)) = -\exp(u_i(t)) \cdot \exp\{-\exp(u_i(t))\}$ and $\phi_i'(u_i(t)) = \{\exp(-u_i(t)^2/2)\}/\sqrt{2\pi}$, respectively. The two functions are non-constant, monotonically increasing and bounded above by 1 and below by 0. Moreover, the two functions are differentiable functions. Thus, complementary log-log and probit functions satisfy the requirements of the UAT [3, 5] for being activation functions. In the output layer, $o_k(t)$, where k is the number of output nodes assumes the linear form. A single hidden layer is sufficient for a MLP to uniformly approximate any continuous function with support in a unit hypercube [2, 3].

The Monte Carlo simulations were performed with 1,000 replications, at the end of the experiments, the average and standard deviation were calculated for the MSE. The simulated data were fitted with different known activation functions known – logit and hyperbolic tangent; and the new activation functions complementary log-log and probit. For the majority of the settings used, the mean values of the measures of error revealed statistically significant differences. The results reveal that the difference in the average MSE of the functions was lower and statistically significant when the reference function was equal to the activation function used in the MLP network. The complementary log-log and probit as activation functions generally presented a lower average MSE than the logit and hyperbolic tangent functions.

References

[1] Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag: New York, 2006.

[2] Cybenko, G. Continuous valued neural networks with two hidden layers are sufficient. Technical report, Department of Computer Science, Tufts University, Medford, MA, 1988.

[3] Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2, 303–314, 1989.

[4] Hartman, E., Keeler, J. D. and Kowalski, J. M. Layered neural networks with gaussian hidden units as universal approximations. *Neural Comput.*, 2(2), 210–215, 1990.

[5] Haykin, S. *Neural Networks: A Comprehensive Foundation,* 2nd ed. Prentice Hall: New Jersey, 2001.

[6] Leung, H. and Haykin, S. Rational function neural network. *Neural Computation*, 5(6), 928–938, 1993.

[7] McCullagh, P. and Nelder, J. A. *Generalized Linear Models,* 2nd ed. Chapman and Hall: London, 1989.

[8] Nelder, J. A. and Wedderburn, W. M. Generalized linear models. *Journal of The Royal Statistical Society*, 3, 370–384, 1972.

[9] Singh, Y. and Chandra, P. A class +1 sigmoidal activation functions for FFANNs. *Journal of Economic Dynamics and Control*, 28(1), 183–187, October 2003.

Adaptive Information Extraction from Web Pages by Supervised Wrapper Induction

Rinaldo José de Lima[1], Frederico Luiz Gonçalves Freitas[1], Bernard Espinasse[2]

[1] Centro de Informática Universidade Federal de Pernambuco Recife PE Brazil,
{rjl4,fred}@cin.ufpe.br
[2] LSIS UMR CNRS 6168 Université d'Aix-Marseille Marseille France,
bernard.espinasse@lsis.org

1 Introduction

In this work, we are concerned with Information Extraction (EI) which comprises techniques and algorithms performing two important tasks: identifying the desired, relevant information from semi-structured or non-structured documents and storing it in appropriate formats for future use. Our focus is adaptive IE systems that can be customized for new domains through training using annotated *corpora* as input. Particularly, we look into automatic *wrapper induction* and Natural Language Processing (NLP) techniques for extraction that rest on the exploitation of structural and grammatical regularities present in documents. Wrappers are procedures to extract data from information resources. Wrapper induction is a technique that uses machine learning algorithms for automatically construct wrappers from a previously annotated corpus [4]. The wrappers we developed are based on the Boosted Wrapper Induction (BWI) algorithm [1] and integrate IE system TIES (Trainable Information Extraction System)[6], developed at ITC-irst. BWI uses the *AdaBoost* algorithm that works by continually reweighting the training examples, and using a base learner (called weak learner) to learn a new classifier repeatedly, stopping after a fixed number os iterations. The classifiers learned are then combined by weighted voting [5]. TIES incorporates the BWI algorithm and automatically induces wrappers from a set of documents annotated with XML tags that identify instances of entities to be extracted. Kauchak [3] has investigated how boosting contributes to the success of the BWI algorithm and studied its performance in the challenging direction of using it as an IE method for unstructured natural language documents. This fact motivated us to extend TIES current version to include Parts-of-Speech (POS) tagging in its preprocessing phase using a POS tagger. Moreover, we have included a module for cleaning up ill-formed tags and attributes of Web pages to produce well-formed XHTML documents which are submitted for tokenisation in TIES architecture.

2 Experiments and Results

The following tables show the first results obtained using the newly extended TIES system on standard tasks for adaptive IE: the CMU Seminars and Austin Jobs announcements and Call for Papers (Pascal Challenge) [2]. We measured the performance in terms of the classical measures in IE domain: *precision*, *recall*, and *F-measure*. We also conducted experiments using various combinations of features in order to systemantically examine their effects on the perfomance of the learning algorithm based on supervised classification.

Table 1: Results without POS Information

Corpus	Precision	Recall	F1-Measure
Seminars	0.974	0.953	0.963
Jobs	0.945	0.778	0.853
CFP	0.891	0.571	0.696

Table 2: Results with POS Information

Corpus	Precision	Recall	F1-Measure
Seminars	0.971	0.964	0.967
Jobs	0.939	0.780	0.853
CFP	0.896	0.591	0.712

3 Conclusion and Future Work

The results we obtained allow us to conclude that the newly extended TIES system, an adaptive IE system based on supervised wrapper induction is comparable with other state-of-the-art IE systems on traditional IE tasks.

Future work includes the improvement of TIES architecture by including new supervised machine learning algorithms, such as Support Vector Machines and C4.5 as learning components for new IE wrappers. Another main issue that is left for future work is related to the tokenizer and feature extraction modules in which we intend to perform the following NLP subtasks, i.e, Named Entity Recognition and Chunking Analysys (for English) and POS tagging for Portuguese and French languages.

References

[1] Freitag, D., Kushmerick, N.: Boosted Wrapper Induction. In: Proceedings of the Seventeenth National Conference on Artificial Intelligence. AAAI-2000,(2000).

[2] Ireson, N., Ciravegna, F.: Pascal Challenge: The Evaluation of Machine Learning for Information Extraction. Machine Learning for the Semantic Web. Dagstuhl Seminar,Dagstuhl, DE (2005).

[3] Kauchak, D., Smarr, J., Elkan, C.: Sources of Success for Information Extraction Methods. Technical Report CS2002-0696. Department of Computer Science and Engineering. University of California, San Diego, January (2002).

[4] Kushmerick, N.: Wrapper Induction: Efficiency and Expressiveness. Artificial Intelligence. vol. 118, (2000).

[5] Shapire, R. E.: A Brief Introduction to Boosting. In Proceedings of the 16th International Joint Conference on Artificial Inteligence (IJCAI) (1999).

[6] TIES. Trainable Information Extraction System. http://tcc.itc.it/research/textec/tools-resources/ties.html, (2004).

Supervised Classification and AUC

Ndèye Niang[1], Gilbert Saporta[1]

Chaire de Statistique Appliquée & CEDRIC
CNAM, 292 rue Saint Martin, 75141 Paris Cedex 03, France

Abstract In supervised classification, ROC curves and AUC are commonly used to evaluate and to compare models performances. Evaluations of AUC are usually done on one validation (hold-out) set. Resampling procedures allow a better use of ROC curves and AUC for predictive purposes.

Keywords: ROC curve, AUC, resampling

1 Measures of performance: Roc curve, Lift curve and Gini index

We focus on supervised classification into two groups. Error rate estimation corresponds to the case where one applies a strict decision rule. But in many other applications one just uses a "score" S as a rating of the risk to be a member of one group, and any monotonic increasing transformation of S is also a score. Usual scores are obtained with linear classifiers (eg Fisher's discriminant analysis, logistic regression) but since the probability $P(G_1|\text{x})$ of classifying an observation $\mathbf{x}$ in the group G1 is also a score ranging from 0 to 1, almost any technique gives a score.

The Receiver Operating Characteristic (ROC) curve synthesizes the performance of a score for any threshold s such that if $S(\mathbf{x}) > s$ then $\mathbf{x}$ is classified in G1. Using s as a parameter, the ROC curve links the true positive rate (or specificity) to the false positive rate (or 1- sensitivity). One of the main properties of the ROC curve is that it is invariant with respect to any increasing (not only linear) transformations of S . Since the ideal curve is the one which sticks to the edges of the unit square, the favourite measure of performance is given by the area under the ROC curve (AUC). Theoretical AUC is equal to the probability of "concordance" : $AUC = P(X_1 > X_2)$ when one draws at random two observations independently from both groups $AUC = \int_{s=+\infty}^{s=-\infty} (1-\beta(s))d\alpha(s)$ where $1-\beta$ is the power of the procedure, and α is the probability of the first kind error. The diagonal corresponds to the worst case where score distributions are identical for both groups. Some practitioners use the lift curve and the area under it (AUL) instead of AUC. The lift curve links the true positive rate (or specificity) to $P(S > s)$. AUC and AUL are linked through the Gini index G : G is the area between the lift curve and the diagonal divided by the area between the ideal lift curve and the diagonal and also twice the area between the ROC curve and the diagonal: $G = 2AUC - 1$.

2 Evaluation of AUC

Let us consider two samples of n_1 and n_2 observations drawn from both groups and some score function S related to the probability of belonging to group 1. A pair of observations x_1 and x_2, one from each group is said to be concordant if $S(x_1) > S(x_2)$. A non parametric estimate of AUC is thus given by the proportion of concordant pairs. The number of concordant pairs is equal to the well known Mann-Whitney's U statistic. Using the relationship between the U statistic and the Wilcoxon W statistic for group1: $W = U{+}n_1(n_1 + 1)/2$. Hanley et al. [1] obtained the standard error of the empirical AUC as :

$$SE = \sqrt{(A(1 - A) + (n_1 - 1)(Q_1 - A^2) + (n_2 - 1)(Q_2 - A^2))/n_1 n_2}$$

where A is the true or theoretical AUC, an unbiased estimates of which being the empirical AUC, $Q_1 = A/(2-A)$ and $Q_2 = 2A^2/(1+A)$. The question is to know how the model will perform for future data (the generalization capacity), provided that future data will be drawn from the same distribution. Evaluating models on the basis of the learning sample may be misleading. If we want to predict capabilities of a method, it is necessary to do so with independent data : it is generally advised to divide randomly the total sample into two parts : the training set and the validation set according to a stratified sampling scheme (the strata are the two groups) without replacement of eg 70% for the training sample and 30% for the validation sample. However in order to avoid a too specific pattern, this random split should be repeated. The performance of the method can then be measured by the AUC computed for all the validation samples : the empirical mean and standard error give an unbiased estimation of future AUC and its standard error and therefore asymptotic confidence interval can be derived.

3 A case study

We exemplify the notions evocated in the previous section on a diabetis data set (http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm). We applied two standard classification techniques : Fisher's linear discriminant analysis (LDA) and logistic regression leading both to a score function. Evaluation of their performances is done by computing AUC for thirty validation sets. The results show the variability of ROC curves which may have very specific and unexpected patterns [3]. It is also shown that AUC has a small but non neglectable variability, average AUC for both methods are lower than AUC computed on the total sample but are unbiased, and LDA performs as well as logistic regression.

References

[1] Hanley, J.A. and McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, **142** (1982) 29-36.

[2] Hanley. J.A, and McNeil, B.J. : A method of comparing the areas under receiver operating characteristic (ROC) curves derived from the same cases. Radiology, **148**(1983) 839-843.

[3] Saporta, G. and Niang, N.: Resampling ROC curves. In IASC meeting on Statistics for Data Mining, Learning and Knowledge Extraction, IASC07 August 30-September 1, Aveiro, Portugal, (2007).

Clustering Electric Load Curves: The Brazilian experience

José Francisco Moreira Pessanha[1], Luiz da Costa Laurencel[2]

[1] Centro de Pesquisas de Energia Elétrica (Cepel), francisc@cepel.br
[2] Universidade do Estado do Rio de Janeiro (UERJ/UFF), getlcl@vm.uff.br

Abstract This contribution presents a brief history of softwares for identifying of typical daily load profiles developed in the Brazilian electric power sector since the eighties. In order to tell this history we present the features of three softwares that have been used by the Brazilian electric distribution utilities, all of then with data mining techniques for clustering.

Keywords: cluster analysis, electric load curves, Brazilian power sector

1 Introduction

The Brazilian electric power sector adopts tariffs based on marginal cost pricing since 1980 [1]. The electricity tariffs are calculated by a methodology, whose origin is the Electricité de France (EDF) and the French 'marginaliste' economists like Allais and Boiteux [2]. In this methodology, an important step is the identification of a few typical daily load profiles from a set of electric load curves measured on a sample of customers. These profiles represent patterns of energy use at different class of customers e.g. residential, commercial, industrial, rural, public lighting, public administration etc. The input data required to the modelling of typical load profiles is a sample of electric load curve measurements. In general, the load curves measures cover a period of two weeks (15 days), where the demand is recorded every 15 minutes by recording meters installed at each consumer point in the sample. In order to reduce the data dimensionality three representative days must be selected from each load curve measurement file: a workday, a Saturday and a Sunday. Each representative daily curve is a vector with 96 points (demand is recorded every 15 minutes). The standard way to identify the typical load profile from a sample of load curves is to perform a clustering of the representative workday load curves classified in a same customer class. The centroid of each cluster defines a typical load profile.

2 Softwares

In the SNACC, the first program developed in the Brazilian power sector in 1982, were adopted the methods of cluster analysis programmed in two computational routines in FORTRAN brought from France: NUDYC and DESCR2. These routines are executed sequentially: first the typical workday load curves are clustered by the nuées dynamique" [1,3,4], a non-hierarchical method programmed in the NUDYC routine, then the clusters centroid vectors are clustered by the Ward method [1,3] (hierarchical method) programmed in the DESCR2 routine. Despite these sophisticated routines to identify the typical load profiles, the SNACC has not a friendly user interface. The program does not show graphs of load curve measures and typical load profiles, an important output for the load curve analysis. This deficiency was the main critical step for the computation of distribution tariff. In order to overcome this deficiency, in 1998 the Brazilian Electric Power Research Center (Cepel) developed the TARDIST system [5] for calculating the distribution tariffs based on marginal cost. This software also has a module to build typical load profiles with a friendly user interface, where, for example, one can choice the three representative days of each load curve measurements based on a graphical output. The software employs only the Ward method in order to derive typical load profiles from a set of workday load curves. The graphical interface shows the workday load curves classified in each cluster and the resulting centroid, i.e., the typical load profile. Also are presented the within and between group sum of squares (in absolute and relative values) for different aggregation level. These statistics and the graphical outputs help us determine the number of typical load profiles. More recently the Cepel developed the ANATIPO, a new software to identify typical load profiles from a sample of load curves measurements. This software incorporates several advances in graphical interface techniques, for example, facilitate the selection of the three representative load curves from a file measurement and also allow move workday load curves from a cluster to another. The software offers three methods of cluster analysis: k-Means, Ward and Fuzzy cluster method (FCM). The typical load profiles obtained by the ANATIPO are organized in electronic worksheets ready to attach in the tariff proposal to be send to ANEEL (the Brazilian regulatory agency). Following the international trend [6,7], nowadays in Brazil the researches on the area have been concentrated on the application of the Self-Organizing Map, a unsupervised neural network.

References

[1] BRASIL, Ministério das Minas e Energia, DNAEE, Eletrobrás, Empresas Concessionárias de Energia Elétrica, Nova Tarifa de Energia Elétrica: metodologia e aplicação, DNAEE, Brasília, 1985.

[2] Boiteux, M. La tarification dês demandes en pointe: application de la théorie de la vente au coût marginal, Revue générale de l'electricité, 1949.

[3] Lebart, L.; Piron, M.; Morineau, A. Statistique exploratoire multidimensionnelle, 3e édition, DUNOD, Paris, 2000.

[4] Bouroche J.M. et Saporta G., L'analyse dês données, PUF, 9e édition, Paris, 2005.

[5] Pessanha, J.F.M., Huang, J.L.C., Pereira, L.A.C., Passos Júnior, R., Castellani, V.L.O. Metodologia e sistema computacional para cálculo das tarifas de uso dos sistemas de distribuição, XXXVI SBPO, São João del Rey - MG,2004.

[6] Debrégeas, A., Hébrail, G. Interactive interpretation of Kohonen maps applied to curves, International Conference on Knowlodge Discovery and Data Mining, New York, August, 1998.

[7] Hébrail, G. Practical data mining in a large utility company, Revue Questiio (Quaderns d'Estadistica i Investigacio Operativa), Vol.25, N.3, pp.509-520, 2001.

Eliciting GAI preference models with binary attributes aided by association rule mining

Sergio Queiroz[1]

Centro de Informática, Universidade Federal de Pernambuco, Av. Prof. Luiz Freire - s/n - Cidade Universitária - CEP 50740-540 - Recife (PE) - Brazil. E-mail: srmq@cin.ufpe.br

Abstract Generalized additive independent (GAI) preference models are, in many situations, more interesting than additive models, as GAI models allow interdependencies between attributes. However, they are more difficult to construct (elicit), not only because the number of questions needed to be asked increases, but also because we must know what the interdependencies are, i.e. the structure of the model. We introduce the use of association rules to select attributes and detect simple interactions between them during an interactive preference elicitation process that intends to build a GAI utility function for the preferences of the user of a recommender system. Using this strategy, we have built a recommender system prototype that suggests touristic sites in a city. We show that, after the elicitation, the recommendation problem can be solved as an instance of the non-linear generalized additive knapsack problem.

Keywords: Preference Elicitation, Multiattribute Recommender Systems, Generalized Additive Preferences, Non-Linear GAI Knapsack Problem.

1 Introduction

The development of decision support systems and web recommender systems has stressed the need for models that can handle users preferences and perform preference-based recommendation tasks. In some applications, like many recommender systems, the decision maker (DM) express his opinions about some alternatives (in a explicit or implicit manner) and the application will try to find among the available objects in a database the ones that are more probable to be liked by him. However, these approaches are only possible when the objects are available in a standardized form, as for example CDs, books, and DVDs [6]. When the objects are composed by configurable attributes, the enumeration of all possible configurations may be prohibitive. This condition characterizes the problems where the space of alternatives has a combinatorial structure, i.e. it is a Cartesian product of variables. In this case, an efficient form of representing and reasoning with preferences over combinatorial domains is needed.

In this way, several current works in preference modeling and decision theory aim at developing compact preference models that achieve a good trade-off between two conflicting goals: i) the power to describe sophisticated decision behaviors; and ii) the practical necessity of keeping the elicitation effort at an admissible level as well as the need for fast procedures to solve preference-based optimization problems.

A good compromise between model simplicity and representational power is reached by GAI (*generalized additive independence*) models [4, 2]. Having the DM preferences modeled as GAI functions brings many benefits. The descriptive power of such models allows interactions between attributes while preserving decomposability (provided that the preferences exhibit some structure). In this way, preferences may be compactly stored. Moreover, given an utility function, we can easily compute the utility of an item, allowing us to quickly calculate its value. But still, there is no easy way to elicit GAI functions.

In this work we use association rule mining [1] in order to select the attributes to expose to the DM at each moment in the elicitation process. Also, association rules are used as a heuristic to detect dependencies between attributes. Using this strategy, we have built a recommender system prototype that suggests touristic sites to visit in a city, under the constraint of the time that the DM will stay at the city. We show that, after the elicitation process, the best recommendations can be obtained as a solution to a non-linear GAI-decomposable knapsack problem. Formally, this correspond to solve this optimization problem:

$$\begin{aligned} \text{maximize} \quad & u(x) \\ \text{under the constraint that} \quad & \sum_{j=1}^{n} w_j x_j \leq c\,, \\ & x = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n. \end{aligned} \tag{1}$$

We suppose that all coefficients w_j, as well as c are non-negative real numbers and that $u(x)$ is a GAI-decomposable utility function over the combinatorial domain of alternatives, i.e. $u : \{0,1\}^n \mapsto \mathbb{R}$. As the utility function $u(x)$ is not additively decomposable, the resulting knapsack problem is much more difficult to solve than for the additive case, given that we cannot exploit the structural independencies to develop more efficient algorithms [3]. We use GAI-Networks [5], a graphical model for the representation of GAI models, in order to efficiently treat the optimization problem.

References

[1] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD*, pages 207–216, Washington, D.C., 26–28 May 1993.

[2] Fahiem Bacchus and Adam J. Grove. Graphical models for preference and utility. In *UAI*, pages 3–10, 1995.

[3] Kurt M. Bretthauer and Bala Shetty. The nonlinear knapsack problem – algorithms and applications. *European Journal of Operational Research*, 138(3):459–472, 2002.

[4] P. C. Fishburn. Interdependence and additivity in multivariate, unidimensional expected utility theory. *International Economic Review*, 8:335–342, 1967.

[5] Christophe Gonzales, Patrice Perny, and Sergio Queiroz. GAI-networks: Optimization, ranking and collective choice in combinatorial domains. *Foundations of computing and decision sciences*, 32(4):3–24, 2008.

[6] Francesco Ricci, Adriano Venturini, Dario Cavada, Nader Mirzadeh, Dennis Blaas, and Marisa Nones. Product recommendation with interactive query management and twofold similarity. In *ICCBR 2003*, volume 2689 of *Lecture Notes in Computer Science*, pages 479–493. Springer, 2003.

Holt's exponential smoothing model for interval-valued time series

André Luis Santiago Maia[1], Francisco de A.T. de Carvalho[1]

Centro de Informática, Universidade Federal de Pernambuco, Av. Prof. Luiz Freire - s/n - Cidade Universitária - CEP 50740-540 - Recife (PE) - Brazil

Abstract Interval-valued time series are interval-valued data that are collected in a chronological sequence through time. This paper adapts an approach to forecasting interval valued-time series based on Holt's exponential smoothing method. In the adapted Holt's method for interval-valued time series, the smoothing parameters are estimated by using techniques for non-linear optimization problems with bound constraints. The practicality of the method is demonstrated by simulation studies and applications using real interval-valued stock market time series.

Keywords: Symbolic Data Analysis, Time Series Forecasting, Interval-Valued Data, Exponential Smoothing.

1 Introduction

Exponential smoothing (ES) methods (e.g., Gardner [5]) have become very popular because of their (relative) simplicity compared to their good overall performance. The Holt's smoothing method (originally presented in Holt [6], reprinted 2004 and Winters [8]), also referred to as double exponential smoothing, is an extension of ES designed for trended time series.

This paper addresses the forecasting of time series with interval-valued data, i.e., when interval-valued variables are collected in an ordered sequence over time, we say that we have an *interval-valued time series* (see Maia et al.
[7]). Thus, an ITS is as

$$\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_n \qquad \text{or} \qquad \begin{bmatrix} X_1^U \\ X_1^L \end{bmatrix}, \begin{bmatrix} X_2^U \\ X_2^L \end{bmatrix}, \ldots, \begin{bmatrix} X_n^U \\ X_n^L \end{bmatrix},$$

where n denotes the number of intervals of the time series (sample size). Interval-valued data arise quite naturally in many situations where such data represent uncertainty (for instance, confidence intervals), variability (minimum and maximum of daily temperature), etc. Interval-valued data have been considered in the field of *Symbolic Data Analysis* (SDA) [2], [1], [4]. Concerning ITS, Maia et al. [7] have introduced approaches to modelling and forecasting based on the ARIMA models, based on an artificial neural networks (ANN) model and based on a hybrid methodology that combines both ARIMA and ANN models.

The interval Holt's exponential smoothing method (Holt$^{\mathrm{I}}$) follows similar representation for usual quantitative data and has the following form:

$$\begin{aligned} \widehat{\mathbf{L}}_t^{\mathrm{I}} &= \mathcal{A}\mathbf{I}_t + (\mathbf{I} - \mathcal{A})(\widehat{\mathbf{L}}_{t-1}^{\mathrm{I}} + \widehat{\mathbf{T}}_{t-1}^{\mathrm{I}}), & (1) \\ \widehat{\mathbf{T}}_t^{\mathrm{I}} &= \mathcal{B}(\widehat{\mathbf{L}}_t^{\mathrm{I}} - \widehat{\mathbf{L}}_{t-1}^{\mathrm{I}}) + (\mathbf{I} - \mathcal{B})\widehat{\mathbf{T}}_{t-1}^{\mathrm{I}}, & (2) \end{aligned}$$

where $\mathcal{A}$ and $\mathcal{B}$ denote the (2×2) smoothing parameters matrices and $\mathbf{I}$ is an (2×2) identity matrix. The estimation of the optimum $\mathcal{A}$ and $\mathcal{B}$ matrices is obtained by the L-BFGS-B (*limited memory algorithm for bound constrained optimization*), method developed by Byrd et al. [3].

The advantage of the use of the model introduced in order to forecast ITS is illustrated through a comparison between the performance of the Holt^{I} (fitting simultaneously the interval boundaries) and the performanc of the usual Holt's exponential smoothing method (fitted independently, on the lower and upper boundaries). The results of the simulation and the applications demonstrated that the usual Holt model performs more poorly than the Holt^{I} model introduced in this paper, i.e., fitting simultaneously the interval boundaries is itself an advantage.

References

[1] L. Billard, E. Diday, *Symbolic Data Analysis. Conceptual Statistics and Data Mining* (Wiley, Chichester, 2006).

[2] H.-H. Bock and E. Diday, *Analysis of Symbolic Data*, Springer, Berlin Heidelberg, 2000.

[3] R. H. Byrd, P. Lu, J. Nocedal and C. Zhu, A limited memory algorithm for bound constrained optimization, *SIAM Journal Scientific Computing*, 16:1190–1208, 1995.

[4] E. Diday, M. Noirhomme-Fraiture, *Symbolic Data Analysis and the Sodas Software* (Wiley, Chichester, 2008).

[5] E. S. Gardner Jr, Exponential smoothing: The state of the art, *Journal of Forecasting*, 4:1–28, 1985.

[6] C. C. Holt, Forecasting seasonals and trends by exponentially weighted averages. *ONR Research Memorandum 52*, Carnegie Institute of Technology, Pittsburgh, Pennsylvania. Reprinted with discussion in *International Journal of Forecasting*, 20, 5–13, 2004.

[7] A. L. S. Maia, F. A. T. De Carvalho and T. B. Ludermir, Forecasting models for interval-valued time series. *Neurocomputing*, 71, 3344–3352, 2008.

[8] P. R. Winters, Forecasting sales by exponentially weighted moving averages. *Management Science*, 6, 324–342, 1960.

Cartão de Pagamento do Governo Federal: uma Análise de Regras de Associação

Dr. Michel Silva, Me. Marcelo Stopanovski, Me. Henrique Rocha, Bel. David Cosac

Secretaria de Prevenção da Corrupção e Informações Estratégicas, Controladoria-Geral da União, SAS, Quadra 1, Bloco A, Edifício Darcy Ribeiro, Brasília/DF, CEP 70070-905

Abstract Desde 2003, o Governo Federal tem trabalhado fortemente pela transparência dos gastos públicos. O Observatório da Despesa Pública e o Portal da Transparência, iniciativas da Controladoria-Geral da União, são exemplos deste esforço. Centenas de operações com Cartões de Pagamento do Governo Federal são realizadas diariamente e aplicações de técnicas de Data Mining têm permitido identificar operações atípicas no uso desses cartões. Em 2008, as informações disponibilizadas pelo Portal da Transparência desencadearam uma reflexão da sociedade sobre os "Cartões Corporativos". Este trabalho apresenta a Análise de Regras de Associação realizada a partir daquele período.

Keywords: cartões de pagamento do governo federal, data mining, regras de associação, transparência dos gastos públicos.

1 Introdução

A Controladoria-Geral da União (CGU) é o órgão do Governo Federal responsável por assistir direta e imediatamente ao Presidente da República quanto aos assuntos que, no âmbito do Poder Executivo, sejam relativos à defesa do patrimônio público e ao incremento da transparência da gestão, por meio das atividades de controle interno, ouvidoria, auditoria pública, correição, combate e prevenção à corrupção.

Além de ser responsável por fiscalizar e detectar fraudes em relação ao uso do dinheiro público, a CGU também é responsável por desenvolver mecanismos de prevenção à corrupção. Parte dessa atividade é exercida na CGU por meio do Observatório da Despesa Pública (ODP). Para realizar seus projetos e ações, a CGU mantêm convênios e parcerias com órgãos públicos, sociedade civil e organizações não governamentais com o intuito de aprimorar e ampliar os instrumentos e as técnicas de combate e de prevenção à corrupção.

O Portal da Transparência, lançado em novembro de 2004, é um canal pelo qual o cidadão pode acompanhar a execução financeira dos programas de governo, em âmbito federal. Estão disponíveis informações sobre os recursos públicos federais transferidos pelo Governo Federal a estados, municípios e Distrito Federal, para a realização descentralizada das ações do governo, e diretamente ao cidadão, bem como dados sobre os gastos realizados pelo próprio Governo Federal em compras ou contratação de obras e serviços, por exemplo.

Centenas de operações com Cartões de Pagamento do Governo Federal (CPGF) são realizadas diariamente e o extrato das transações é disponibilizado nesse portal mensalmente. Aplicações de técnicas de Data Mining têm permitido identificar operações atípicas no uso desses cartões. Em 2008, as informações disponibilizadas pelo portal desencadearam uma atenção especial da sociedade sobre o assunto. Este trabalho apresenta a Análise de Regras de Associação realizada a partir daquele período.

2 Análise de Regras de Associação e o Escândalo dos "Cartões Corporativos"

A CGU, para um trabalho piloto visando a criação do ODP, efetuou em 2008 análise e processamento de dados em busca de padrões presentes nas transações financeiras referentes ao uso dos CPGF. Após o recebimento das informações, foi estabelecido que os dados a serem minerados envolveriam apenas os gastos não-sigilosos, por permitirem o uso de variáveis mais significativas, o que correspondia à época a 225.892 transações de um total de 237.189, o equivalente a aproximadamente 95% do total das transações efetuadas no ano de 2007, objeto da análise. O volume total de recursos envolvidos em transações foi de R$ 59.681.032,74.

O processo de mineração de dados envolveu o uso da técnica de regras de associação em dados categóricos para as variáveis: órgão superior do portador do cartão de crédito, subclasse do estabelecimento comercial (compõe o CNAE Fiscal, extraído das bases de dados da Receita Federal do Brasil) e faixa de valor da despesa.

O suporte estabelecido para a mineração de dados usando a técnica de regras de associação foi de 0,1% e a confiança de 50%. Esses parâmetros foram obtidos a partir do ajuste fino ocorrido em duas iterações de processamento, dado o grande volume de transações.

É válido ressaltar que o objetivo desse trabalho foi apenas planejar e executar processos de mineração de dados em busca de padrões, de maneira metodológica, e com o uso de técnicas amplamente conhecidas. Como o próprio processo de mineração de dados preconiza, cada uma das regras de associação encontradas deve ser objeto de análise individual pelo especialista de domínio para que sejam verificadas sua relevância e utilidade em prol da atividade de auditoria, fiscalização e controle dos gastos públicos, missão desta Controladoria. Este trabalho de análise pelos especialistas gerou um monitoramento dos padrões detectados. Segue abaixo dois exemplos de regras encontradas:

I) {"LOCAÇÃO DE AUTOMÓVEL SEM CONDUTOR", "R$ 1000 a R$ 1500"} ⟶ {"SEC. ESP. DE POLÍTICAS DE PROMOÇÃO DA IGUALDADE RACIAL"}: suporte = 1,15% e confiança = 86,67%. Ressalta-se que esse exemplo é exatamente o pivô das discussões sobre cartões, configurado no aluguel de carros em viagens.

II) {"COMÉRCIO VAREJISTA DE COMBUSTÍVEIS PARA VEÍCULOS AUTOMOTORES", "IBGE"} ⟶ {"R$ 50 a R$ 100 "}: suporte = 5,11% e confiança = 52,53%.

3 Comentários Finais

Desta forma, o presente trabalho levantou um total de 155 regras de associação, que permitiram a descoberta de várias anomalias. A análise de regras de associação se mostrou bastante útil para o direcionamento dos trabalhos de auditoria.

Hoje o ODP da CGU mantém um quadro de indicadores preventivo sobre os padrões identificados. Tal quadro é muito semelhante ao utilizado em grandes empresas de cartões privados. O trabalho gerou ainda lampejos de predição, permitindo análises antecipadas que depois conformaram o interesse social e da mídia sobre o assunto.

Beyond the non-probabilistic symbolic regression models for interval variables

Eufrásio de A. Lima Neto[1], Gauss M. Cordeiro[2], Francisco de A.T. de Carvalho[3]

[1] Departamento de Estatística, Universidade Federal da Paraíba - Cidade Universitária s/n - CEP 58051-900 - João Pessoa (PB) - Brazil

[2] Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - Dois Irmãos - CEP 52171-900 - Recife (PE) - Brazil

[3] Centro de Informática, Universidade Federal de Pernambuco, Av. Prof. Luiz Freire - s/n - Cidade Universitária - CEP 50740-540 - Recife (PE) - Brazil

Abstract This paper presents a overview about the symbolic regression models to interval-valued data. The major symbolic regression methods proposed in literature visualized the problem like a optimization point of view. Lima Neto et. al. (2009) proposed a new symbolic regression model for interval variables, called bivariate generalized linear model (BGLM), which are based on bivariate exponential family of distributions [5], making possible the use of statistical inference techniques and goodness-of-fit measures over symbolic regression models.

Keywords: Symbolic Regression Models, Bivariate Generalized Linear Models, Interval-valued Data

1 Introduction

In regression analysis of quantitative data, the items are usually represented as a vector of quantitative measurements [7]. However, due to recent advances in information technologies, it is now common to record interval-valued data. In the Symbolic Data Analysis (SDA) framework [1, 3, 4], interval-valued data appear when the observed values of the variables are intervals from the set of real numbers $\Re$. Moreover, interval-valued data arise in practical situations such as recording monthly interval temperatures in meteorological stations, daily interval stock prices, among others. Another source of interval data is the aggregation of huge data-bases into a reduced number of groups, the properties of which are described by symbolic interval variables.

Billard and Diday [2] presented the first approach to fit a linear regression model to a symbolic interval-valued data set. Their approach consists of fitting a linear regression model to the midpoint of the interval values assumed by the variables in the learning set and then applies this model to the lower and upper limits of the interval values of the explanatory variables to predict, respectively, the lower and upper limits of the interval values of the dependent variable. Lima Neto and De Carvalho [8] improved the former approach presenting a new method based on two linear regression models, the first regression model over the midpoints of the intervals and the second one over the ranges, which reconstruct the bounds of the interval-values of the dependent variable in a more efficient way.

Despite recent contributions to symbolic regression models, current approaches view the problem from an optimization point of view and do not consider the probabilistic aspects related to regression models. This make it impossible to use inference techniques over the parameters estimates, such as hypothesis tests or confidence intervals.

Generalized linear models represent a major synthesis of regression models by allowing a wide range of types of response data and explanatory variables to be handled in a single unifying framework. These models are based on the exponential family of distributions and represent a very important regression tool due to their flexibility and applicability in practical situations [6]. Iwasaki and Tsubaki [5] introduced a class of bivariate generalized linear models (BGLMs) based on the bivariate exponential family of distributions with an application to meteorological data analysis.

Lima Neto et. al. (2009) considered the BGLM as an important tool for solving problems related to SDA and presented a model based on bivariate Gaussian distribution. They also presented an alternative way to estimate the dispersion parameter ϕ and the coefficient of correlation ρ. The latter is based on the log-likelihood profile method. Additionally, the goodness-of-fit measures, which are not addressed by Iwasaki and Tsubaki, were considered by them. Application to a real interval data sets demonstrated that the BGLM method presented a better fit when compared with the non-probabilistic symbolic regression methods proposed by [2] and [8]. However, the authors recommend a simulated study in future works for a more consistent conclusion about the BGLM method.

References

[1] Book, H.H., Diday, E.: Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data. Springer-Verlag (2000).

[2] Billard, L., Diday, E.: Regression Analysis for Interval-Valued Data. Proceedings of the Seventh Conference of the International Federation of Classification Societies. Springer-Verlag (2000) 369-374.

[3] Billard, L., Diday, E.: Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley (2006).

[4] Diday, E., Fraiture-Noirhomme, M.: Symbolic Data Analysis and the SODAS Software. Wiley-Interscience (2008).

[5] Iwasaki, M., Tsubaki, H.: A bivariate generalized linear model with an application to meteorological data analysis. Statistical Methodology **2** (2005) 175-179.

[6] McCullagh, P., Nelder, J.: Generalized Linear Models. Chapman & Hall (1989).

[7] Montgomery, D.C., Peck, E.A.: Introduction to Linear Regression Analysis. John Wiley (1982).

[8] Lima Neto, E.A., De Carvalho, F.A.T.: Centre and range method to fitting a linear regression model on symbolic interval data. Computational Statistics and Data Analysis **52** 1500–1515.

[9] Lima Neto, E.A., De Carvalho, F.A.T., Cordeiro, G.M, Anjos, U.U., Costa, A.G.: Bivariate Generalized Linear Model for Interval-Valued Variables. Proceedings of the 2009 IEEE International Joint Conferences on Neural Networks. IEEE (2009) accepted for publication.

A Two Stage Clustering Method Combining Self-Organizing Maps and Ant K-means

Jefferson R. Souza, Teresa B. Ludermir and Leandro M. Almeida

Center of Informatics, Federal University of Pernambuco, Av. Prof. Luis Freire, s/n, Cidade Universitária Recife/PE, 50732-970, Brazil
{jrs2, tbl, lma3}@cin.ufpe.br

Abstract This paper proposes a clustering method SOMAK, which is composed by Self-Organizing Maps (SOM) followed by the Ant K-means (AK) algorithm. The aim of this method is not to find an optimal clustering for the data, but to obtain a view about the structure of data clusters. SOM is an Artificial Neural Network, which has one of its characteristics the nonlinear projection. AK is a meta-heuristic approach for solving hard combinatorial optimization problems based on Ant Colony Optimization (ACO). The SOMAK has a good performance when compared with some clustering techniques and reduces the computational time.

Keywords: SOM, ACO and Unsupervised Learning.

With the substantial reduction of cost data storage, a great improvement in the performance of computers and the popularization of computer nets, a great amount of data information is being produced every day everywhere. So, a great quantity scale of databases has created the necessity of developing some techniques of data processing useful for the clustering of data or data mining [1]. The techniques used in this paper were: K-means, SOM, SOM followed by the K-means (SOMK) and SOMAK. K-means is one of the simplest algorithms of non-supervised learning to solve the clustering problem. The aim is to divide the data set within k clusters fixed a priori [4]. AK [5] is a recently proposed meta-heuristic approach for solving hard combinatorial optimization problems named ACO [3].

The method proposed in this paper, SOMAK, can be seen in Fig. 1. SOMAK uses SOM net[1][2] as a classifier of characteristics about the entry data instead of clustering the data directly. First, a large set of prototypes is formed by using SOM. The prototypes can be interpreted as proto-clusters, which are in the next step combined to form the true clusters. For the execution of the experiments were used: synthetic data (I, II, III), real data (IV, V), the method Monte Carlo, to check the efficiency of the clustering methods. The number of clusters and its centroids are obtained by SOM net and then uses AK to find the definite solution. Table 1 shows a comparison between SOMAK and SOMK to obtain a smaller number of clusters. It is important to mention in Table 1 that the fact of the SOMAK method increasing the number of clusters does not mean to say that bad, perhaps this increase may be necessary to have an improvement of entropy. We concluded that the experimental results are statistically independent according to the application of Test t and applied as well for the entropy [7] as for the computational time and with 5% of significance degree it showed that SOMAK is better than SOMK. The

[1] Training of SOM net freely available in the package Matlab SOM Toolbox which was used in the implementation of the proposed method. For further information, see URL http://www.cis.hut.fi/projects/somtoolbox/.

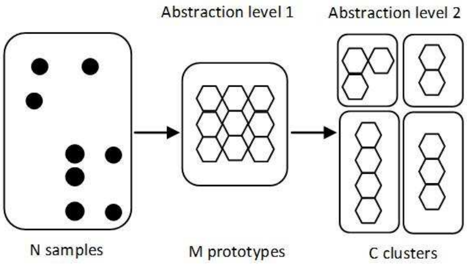


Figure 1: The first level of abstraction is obtained by using SOM. Algorithm SOMAK creates the second level of abstraction carrying out the cluster of M prototypes [6].

Table 1: Results of the size of clusters obtained by the test set

Data sets	Initial Cluster	SOMK	SOMAK
Lines(I)	10	6	3
Banana(II)	2	7	4
Highleyman(III)	2	3	4
CMC(IV)	3	9	4
Glass(V)	6	5	3

benefit of this approach is the reduction of the computational cost. The second advantage is the reduction of the clusters size. The reduction of the noise is another benefit. So, SOMAK is a robust method of clustering and it can be applied to a lot of different kinds clustering problems or combined with some other techniques of data mining to obtain more promising results. For future works, the idea is readjusting the SOMAK algorithm with the purpose of reducing its computational time when compared with the methods described in this paper.

References

[1] Everitt, B. S., Landau, S. and Leese, M.: Cluster Analysis. Edward Arnold. (2001)

[2] Kohonen, T.: The self-organizing map. Neurocomputing. **21** (1998) 1–6

[3] Dorigo, M. and Stützle, T.: The Ant Colony Optimization Metaheuristic: Algorithms, Applications, and Advances. Technical Report IRIDIA (2000)

[4] Mitchell, T.: Machine Learning. McGraw-Hill. (1997) 352p

[5] Kuo, R. J., Wang, H. S., Tung-Lai Hu and Chou, S. H.: Application of Ant K-Means on Clustering Analysis. Computers and Mathematics with Applications. **50** (2005) 1709–1724

[6] Vesanto, J. and Alhoniemi, E.: Clustering of the Self-Organizing Map. IEEE Transactions on Neural Networks. **11** (2000) 586–600

[7] Tan, P., Steinbach, M. and Kumar, V.: Introduction to Data Mining. Pearson. (2006)