Inf Syst E-Bus Manage
DOI 10.1007/s10257-014-0242-7

ORIGINAL ARTICLE

A recommendation engine for disease prediction

Francesco Folino · Clara Pizzuti

Received: 29 April 2013 / Revised: 20 January 2014 / Accepted: 13 February 2014

Abstract An approach for disease prediction that combines clustering, Markov models and association analysis techniques is proposed. Patient medical records are first clustered, and then a Markov model is generated for each cluster to perform predictions about illnesses a patient could likely be affected in the future. However, when the probability of the most likely state in the Markov models is not sufficiently high, the framework resorts to the association analysis. High confidence rules generated by recurring to sequential disease patterns are considered, and items induced by these rules are predicted. Experimental results show that the combination of different mining models gives good predictive accuracy and it is a feasible way to diagnose diseases.

Keywords Disease prediction · Data mining · Association analysis · Clustering · Markov models

1 Introduction

In the last few years we are witnessing an increasing interest in the application of computational science methods for health care information and management systems. Exploitation of information technologies could significantly improve both efficiency and effectiveness of health care strategies, and reveal their importance because of the implications they could have in every day life of individuals.

F. Folino · C. Pizzuti (✉)
National Research Council of Italy—CNR, Institute for High Performance Computing and Networking—ICAR, Via P. Bucci 41C, 87036 Rende, CS, Italy
e-mail: pizzuti@icar.cnr.it

F. Folino
e-mail: ffolino@icar.cnr.it

Published online: 26 February 2014

An emerging viewpoint aims at identifying prospective health care models to determine the risk for individuals to develop specific diseases. In fact, prevention or intervention at the disease's earliest onsets allows advantages for both the patient, in terms of life quality, and the medicare system, in terms of costs. However, recognizing the origin of an illness is not an easy task because it can be generated by multiple causes. Physicians prescribe laboratory tests only after the appearance of patient's symptoms, and use family and health history to assess the hypothesized problem. The approach is thus reactive, i.e., a medical treatment is undertaken only after the patient has already developed the disease, rather than proactive.

However, hospitals and physicians collect thousands of patient clinical histories that include valuable information regarding illness correlations and development. The patient medical records contain important enlightenment regarding the co-occurrences of diseases affecting the same individual. A *comorbidity* relationship between two illnesses exists whenever they appear simultaneously in a patient more than chance [1]. Although comorbidity is very common in the population and its extension increases with age, few investigations have been conducted on patient's comorbidity conditions [2]. The comorbidity relationships between diseases, however, could be exploited to build a model that predicts the diseases a patient could have in the future.

Advanced risk assessment tools are currently at disposal, mainly based on statistical techniques [3, 4]. Another approach for addressing the problem [7], which is gaining increasing interest, is the use of methodologies coming from the fields of knowledge discovery [5].

In this paper we propose a recommendation engine named *CORE* (*COmorbidity-based Recommendation Engine*), that combines techniques coming from the data mining and statistics fields, to determine the risk of an individual to develop future diseases on the base of her/his past patient medical history. A patient is represented by means of a vector of diagnosed disease codes, defined by the *International Classification of Diseases, Ninth Revision, Clinical Modification* ICD-9-CM, and a disease is predicted by exploiting both its medical history and the information regarding other patients having a similar clinical course. The model at the base of the recommendation engine integrates *clustering*, *Markov models*, and *association analysis* with the aim of obtaining specialized and accurate prediction models.

The paper is organized as follows. The next section describes existing approaches to disease prediction. Section 3 introduces the concepts of clustering, Markov models, and association rules. In Sect. 4 the recommendation engine is described. Section 5 shows with an example how the predictions are obtained. In Sect. 6 the data set of medical records used for the experiments is described. Section 7 describes the measures used to assess the quality of the obtained results. Section 8 reports the evaluation of the proposed approach on the patient data set. Section 9 concludes the paper.

2 Related work

Many approaches and tools for the risk assessment of developing illnesses in the medical context have been proposed. These proposals are mainly based on statistical

techniques and use the family history as well as the results of patient clinical tests designed for specific purposes [3, 4].

A general predictive model to assess disease risks has been proposed by Davis et al. [6, 7]. The model is based on patterns of co-occurrences across the medical patient records instead of laboratory tests. Each patient can be associated with the list of diseases she/he has been affected during his life. Groups of illnesses occurring frequently in many patient records can be exploited to capture comorbidity relations and generate predictions about the diseases a patient can incur in, given the past history of his health conditions. A patient is represented by a vector of diagnosed disease ICD-9-CM codes, and a prediction is made on the base of other similar patients. Davis et al. [6, 7] used the patient clinical history, i.e. the diseases a patient was diagnosed for each inpatient hospital visit, to predict the illnesses of the subsequent visits. They built a collaborative assessment and recommendation engine, named *CARE*, that relies on the collaborative filtering methodology [8] for producing recommendations to people, by collecting preferences from users having similar behaviors. In the medical context, users are patients and their behavior corresponds to patient medical history. The predictions on a patient p are done by comparing the individual medical history with the medical histories of a set of patients I, called *training patient set*. The training set is required to contain patients sharing at least two diseases with p. The approach is based on the concept of similarity between the testing patient p and the training patients contained in I. The computation of similarity produces a ranked list of diseases constituting the predictions for the future visits of p. The similarity between two patients is defined by taking into account the random expectation and the inverse frequency of each disease. The inverse frequency is included to reduce the weight of very common sickness since, as the authors note, sharing a rare disease is more informative than, for example, having hypertension. Experiments on a data set of Medicare records of elderly patients showed good prediction accuracy.

Steinhaeuser and Chawla [9] used a hybrid technique based on collaborative filtering and nearest neighbor classification. The similarity between two patients is computed by using the *Jaccard coefficient* [10, 11], which is the normalization of common diseases that two patients have, with respect to their union. Also in this approach, given a patient p, a ranked list of diseases p could develop in the future is computed by considering his k nearest neighbors, i.e. the most similar patients of p. Experimental results on Medicare beneficiaries, aged at least 65 at the time of the first visit, showed that a good percentage of diseases were predicted as expected. A disease network was also built and their structural properties studied.

Analogously to the described approaches, we adopt a representation of patient clinical history based on the ICD-9-CM codes of the diseases the patient has been diagnosed. However, in our case, because of the data set size and characteristics, we have only one patient record storing the overall medical history of that patient, and not a record for each disease diagnosed at a particular hospital visit for that patient. Furthermore, we use a completely different approach of collaborative filtering that combines statistical and machine learning techniques, as it will be described in the next section.

3 Technique description

Let m be the number of patients contained in the data set DS of patient medical histories, and $D = \{d_1, \ldots, d_n\}$ the set of all distinct illnesses appearing in DS. From this data set a new data set $T = \{t_1, \ldots, t_m\}$, where each $t_i \subseteq D$ is a patient medical record of variable size constituted by that sequence of ICD-9-CM disease codes the patient t_i has been affected up to now, is generated. Thus, T represents the medical histories relative to each of the m patients. Table 1 shows an example of a data set T constituted by five patient records, each containing a different number of disease codes.

Before giving the details of the proposed approach, in this section the related concepts necessary to describe it are reported. Thus, in the following, the techniques exploited in our system—specialized for the medical context—are introduced.

3.1 Clustering

Clustering [5] is a well-known data analysis technique that groups similar data objects in clusters such that objects of different groups are dissimilar. Grouping the set T of patients in k groups having similar disease history, can help us to mine more specialized models, tailored for particular groups of patients sharing part of their diseases. In the following, we will describe a clustering method which is a variant of the traditional k-means [12, 13] able to deal with categorical tuples of variable size, like those present in the dataset T.

Let assume k be the number of clusters, this algorithm partitions T into k clusters $C = \{C_1, \ldots, C_k\}$ in a way that high intra-cluster similarity and low inter-cluster similarity are guaranteed. C is a partitioning of T, i.e., $\bigcap_{i=1\ldots k} C_i = \emptyset$ and $\bigcup_{i=1\ldots k} C_i = T$. Each record $t_i \in T$ is assigned to a cluster C_j according to its distance $d(t_i, r_j)$ from a vector r_j that represents the cluster at hand, and it is called the *cluster representative*. Formally, the clustering algorithm finds a partition C such that:

(1) *for each* C_i *the representative* r_i *is computed*;
(2) $t_i \in C_j$ *iff* $d(t_i, r_j) < d(t_i, r_l)$ *for* $1 \leq l \leq k, j \neq l$;
(3) C *minimizes the cost function*:

$$Q_k = \sum_{i=1}^{k} \sum_{t_j \in C_i} d(t_j, r_i) \tag{1}$$

Table 1 Set T of patient records

T	ICD9 codes of medical records
t_1	401, 715, 722, 723
t_2	401, 721, 715, 722, 723
t_3	401, 721, 715, 722
t_4	241, 255, 595, 780
t_5	241, 255, 272, 595, 780

In practice, the algorithm works as follows. Firstly, k records are randomly selected from T. They represent the initial cluster centers. Then, each other $t_i \in T$ is assigned to a cluster on the base of condition (2) above. The algorithm updates the representative of each cluster and re-assigns each record consequently. The iterations terminate when the representatives do not change any more, i.e., the condition (3) holds.

It is worth noting that the schema above is parametric w.r.t. the definitions of both distance d and representative r. Since in our scenario we deal with categorical data, we used a kind of distance that proved to work very well in this setting: the *Jaccard* distance. This measure is derived by the *Jaccard coefficient* [10, 11] which is based on the idea that the similarity between two itemsets is directly proportional to the number of their common items and inversely proportional to the number of different ones. Therefore, given two records t_i and $t_j \in T$, the Jaccard distance can be defined as:

$$d(t_i, t_j) = 1 - \frac{|t_i \cap t_j|}{|t_i \cup t_j|} \tag{2}$$

Another important aspect is the suitable definition of the cluster representative r. Intuitively, the representative should model the content of the cluster in order to make trivial the interpretation of the cluster itself. Among various possibilities, an easy and effective way for building the representative consists in using the frequent items belonging to the cluster itself [12]. The frequency degree can be controlled by introducing a user-defined threshold value γ representing the minimum percentage of occurrences an item must have for being inserted into the cluster representative. More formally, given $T_{C_i} = \{t_1, \ldots, t_q\}$ the set of records belonging to the cluster C_i, $D_{C_i} = \bigcup_i t_i = \{d_1, \ldots, d_p\}$ the set of items of C_i, i.e. the disease codes, and $\gamma \in [0,1]$, then the representative r_{C_i} for the cluster C_i can be computed as follows:

$$r_{C_i} = \{d \in D_{C_i} | f(d, T_{C_i})/q \geq \gamma\} \tag{3}$$

where $f(d, T_{C_i}) = |\{t_i \in T_{C_i} | d \in t_i\}|$ is the number of medical records of cluster C_i in which d appears.

Clearly, the clustering algorithm assumes that the number of clusters k has to be fixed at the beginning. Thus, another open issue is how to set k in order to obtain the best partitioning. Ideally, the best partitioning is achieved for that value k^* in correspondence of which the cost function Q_k has its global minimum. However, finding k^* in this way is an *NP-hard* problem and then it is unfeasible in practice. Therefore, we pragmatically recurred to a sub-optimal solution: we iterated the clustering algorithm by ranging k in $[1,|T|]$ until the first, local minimum for Q_k is reached.

For a better comprehension of the clustering adopted, let us consider the set of medical records in Table 1. Let also assume $k = 2$ be the number of clusters that minimizes the cost function Q_k and $\gamma = 0.5$ be the minimum percentage of occurrences a disease must have for being inserted into the cluster representative (see Eq. 3). On the base of the above parameters, it is easily verifiable that the clustering algorithm finds the clusters C_1 and C_2, as reported in Tables 2 and 3,

respectively. Furthermore, the clusters are equipped with the representatives: $r_{C_1} = \{401, 721, 715, 722, 723\}$ and $r_{C_2} = \{241, 255, 272, 595, 780\}$.

3.2 Markov models

Markov models are a well known technique for understanding stochastic processes and have been extensively used as prediction models because of the good accuracy levels they may reach.

Deshpande and Karypis [14] represent Markov models as a triple $\langle A, S, TPM \rangle$, where A is a set of actions, S is a set of states, and $TPM = \mid S \mid \times \mid A \mid$ is a transition probability matrix, where an entry $tpm_{i,j}$ is the probability that the action j is performed when the process is in the state i. The simplest Markov model, known as *1st-order Markov model*, predicts the next action by looking at only the previous action. In general, a *w-order Markov model* makes predictions by considering the last w actions.

In the context of predicting user's web behavior, Deshpande and Karypis [14] identify the input data for building the Markov models as web sessions, i.e. the sequence of pages accessed by a user during a visit to a specific site. Thus, the actions are the pages of the web site, and the states are the w consecutive web pages observed in different sessions.

In our medical context, the actions are the ICD-9-CM disease codes, the web sessions are the set $T = \{t_1, \ldots, t_m\}$ of patient medical records, and the states are the w consecutive disease codes $\{t_{i1}, \ldots, t_{iw}\}$ observed in T. The transition matrix TPM is then computed by counting how many times the code in position j appears after the state i.

For example, let consider the set of patient medical records reported in Table 4. The 1st-order Transition Probability Matrix (Table 5) is such that each state is constituted by only a disease code, thus there are six different possible states. The matrix entry in position (2, 3) is 3 because there are three medical records (t_2, t_3, t_4) in which the code 715 appears after the state $s_2 = 437$. Then, the probability that the disease 715 will be predicted as next disease after 437 is 1.

Analogously, the 2nd-order TPM (Table 6) contains the couples of codes appearing in sequence. For instance, as for the state {715, 722} and the disease code

Table 2 Cluster C_1 of patient records

T	ICD9 codes of medical records
t_1	401, 715, 722, 723
t_2	401, 721, 715, 722, 723
t_3	401, 721, 715, 722

Table 3 Cluster C_2 of patient records

T	ICD9 codes of medical records
t_4	241, 255, 595, 780
t_5	241, 255, 272, 595, 780

Table 4 A set of four patient medical records

T	ICD9 codes of medical records
t_1	{401, 715, 722, 723}
t_2	{401, 437, 715, 722, 756}
t_3	{401, 437, 715, 722, 723}
t_4	{437, 715, 722, 756}

723, the entry of the matrix in position (4, 5) is 2 because 723 appears twice (t_1 and t_3) after the couple {715, 722}.

Once the TPM of a fixed order is computed, for performing a prediction, given a sequence of ICD-9-CM codes, it is sufficient to look up at the TPM and extract the disease having the highest frequency. For instance, given the set {715, 722}, you may indifferently choose to leverage both the 1st-order and 2nd-order TPM to make the prediction. By resorting to the 1st-order TPM, the next predicted disease after 715 is 722 (see Table 5), whereas if the 2nd-order TPM is used instead, you can equally foresee that the next disease after {715, 722} can be either 723 or 756 (see Table 6).

A main drawback of Markov models is that, in order to obtain good predictive accuracy, higher-order models must be used. However, higher-order models are computing demanding because of the high number of states that can be generated.

3.3 Association analysis

Association analysis [5] is an important data mining methodology for discovering interesting hidden relationships in large data sets. It relies on the concept of *frequent itemset* to extract strong correlations among the items constituting the data set to study. Originally, association analysis has been applied to the market basket data, where each item represents the purchase done by a customer. A transaction is defined as the set of items purchased at the same time by the same customer. However, it can be easily transposed into the medical context by associating an item with a disease, and by considering an itemset, i.e. a transaction, as the set of diseases t_i a single patient had along his life until the present. Groups of diseases occurring frequently together in many transactions are referred to as *frequent itemsets*. The concept of frequency is formalized through the concept of *support*.

Given a set $I = \{I_1, \ldots, I_k\}$ of frequent itemsets on the dataset T, the support of an itemset $I_i \in I$, $supp(I_i)$, is defined as

$$supp(I_i) = \mid \{t \mid t \in T, I_i \subseteq t\} \mid \tag{4}$$

where $\mid . \mid$ denotes the number of elements in a set. Basically, the support determines how often a group of diseases appears together. It is an important measure since very low support discriminates those groups of items occurring only by chance. Thus, a frequent itemset, in order to be considered interesting, must have a support greater than a fixed threshold value, *minsup*.

An association rule is an implication expression of the form $X \Rightarrow Y$, where X and Y are disjoint itemsets. The importance of an association rule is measured by both its

Table 5 1st-order transition probability matrix corresponding to the patient medical records of Table 4

1st order	401	437	715	722	723	756
$s_1 = 401$	0	2	1	0	0	0
$s_2 = 437$	0	0	3	0	0	0
$s_3 = 715$	0	0	0	4	0	0
$s_4 = 722$	0	0	0	0	2	2
$s_5 = 723$	0	0	0	0	0	0
$s_6 = 756$	0	0	0	0	0	0

support and *confidence*. The support σ of a rule determines how often a rule is applicable to a data set and it is computed as

$$\sigma(X \Rightarrow Y) = \frac{supp(X \cup Y)}{\mid T \mid} \tag{5}$$

The confidence τ, instead, determines how frequently items in Y appear in transactions that contain X. It is formally defined as

$$\tau(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \tag{6}$$

Frequent itemsets having a support value above a minimum threshold are used to extract high confidence rules, and can be exploited to build a risk prediction model by matching the medical record of a patient against the patterns discovered by the model. In this scenario, if we consider a rule like $X \Rightarrow \{d\}$, where $X \subseteq D$ is a subset of diseases and d is a single disease, a high support σ determines that the rule is very frequent applicable to the dataset, whereas a high confidence τ allows for reliably inferring that d will appear together with the items contained in X.

Example 1 In order to better understand these concepts, let consider T be the set of the five medical records reported in Table 1, and {401,715} $\Rightarrow$ 723 an association rule mined on it. Since the support *supp* for {401, 715, 723} is 2 and the total number of medical records is 5, thus the rule support σ will be $2/5 = 0.4$. The rule confidence τ is obtained, instead, as $supp(\{401, 715, 723\})/supp(\{401, 715\}) = 2/3 = 0.67$.

In the next section we will show that by combining clustering, low order Markov models and association rules, good values of prediction accuracy, keeping moderate runtime requirements, can be obtained.

4 A framework for disease prediction

A general architecture for the COmorbidity-based Recommendation Engine (*CORE*) is depicted in Fig. 1. *CORE* consists of two main components: an off-line component for the *model generation*, and an on-line component for the *disease prediction*. The model generation component involves a preprocessing step (*Data*

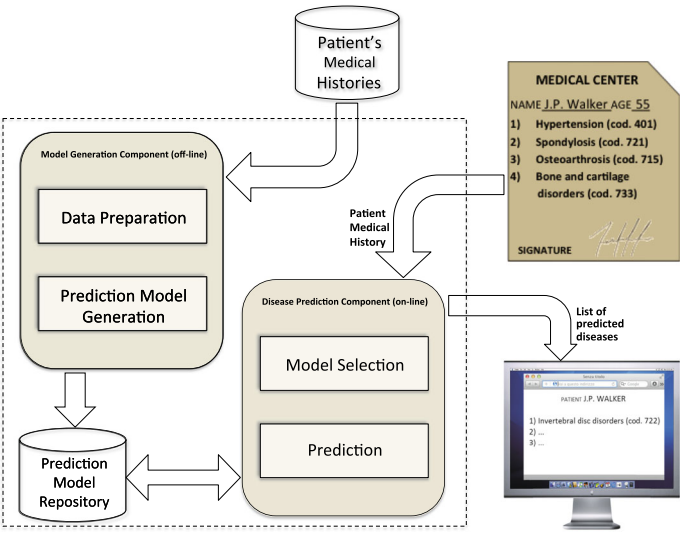

Fig. 1 Architecture of the disease prediction system *CORE*

Preparation) to transform the raw data into a transactional set of patient records constituted by a sequence of ICD-9-CM codes (i.e., the list of diseases a patient had in the past). Then, a *prediction model* is generated, as explained below, in order to fulfill the prediction. The off-line component, instead, given a new patient, first selects a proper model, and then performs predictions on him.

Definition 1 A disease prediction model *DPM* for a dataset T can be defined as a couple $\langle C, M \rangle$, where $C = \{C_1, \ldots C_k\}$ is a clustering of T and $M = \{M_1, \ldots M_k\}$, where each M_i is the prediction model built on top of C_i.

Definition 2 Given a set $I_{C_i} = \{i_1, \ldots, i_l\}$ of frequent itemsets $i_j \in I_{C_i}$ induced on a cluster $C_i = \{t_1, \ldots, t_p\}$, the support $\sigma \in [0,1]$ of a generic itemset i_j w.r.t. the cluster C_i is defined as:

$$\sigma(i_j) = \frac{| \{t_i \mid i_j \subseteq t_i, t_i \in C_i\} |}{| C_i |} \tag{7}$$

where $| \, . \, |$ denotes the number of elements in a certain set.

Given a data set $T = \{t_1, \ldots, t_m\}$ of medical records, and fixed the order w of Markov models, the number k of clusters, the threshold γ for computing the cluster representatives, the support σ and the confidence τ to compute association rules, then the construction of the prediction models performs the following steps:

1. *Cluster medical records in k clusters* $\{C_1, \ldots, C_k\}$;
2. *Build a Markov model* $MM^w_{C_i}$ *of order w for each cluster* C_i;
3. *Compute the association rules for each cluster with support* σ *and confidence* τ.

Once the models are built for each discovered cluster, if a new medical record is given, the prediction phase encompasses three main steps:

1. *Cluster Assignment*, where the patient is recognized as member of a cluster C_i by matching him against each cluster representative;
2. *Model Selection*, where the model M_i (relative to the corresponding cluster) is selected;
3. *Prediction*, where the proper prediction is performed by exploiting M_i.

More in detail, the *prediction* step works in this way. We use a window of size w over the medical records for capturing the patient history depth used for the prediction. The size w means that we apply a Markov model of order w, thus only the last (in time order) w diseases appearing in the record influence the computation of possible forthcoming illnesses. If the Markov model prediction produces either a no state or a state having a not enough high probability (as explained below), the association rules are used instead. In this case, we alternatively resort to the frequent itemsets of size $w + 1$ induced on C_i that contain the w items appearing in the current patient medical record $t_i \in C_i$. The prediction of the next disease is based on the confidence of the corresponding association rule whose antecedent is constituted by the w frequent items of t_i, and the consequent is exactly the disease to be predicted. If this rule has a confidence value greater than a fixed threshold, then its consequent is added to the set of predicted illnesses.

Let us now perform a prediction on a patient medical record $t_i^w = \{t_1^i, \ldots, t_w^i\}$ of size w recognized belonging to the cluster C_i. We first apply the wth-order Markov model learned on the cluster C_i. If t_i^w matches a state in this model, the probability of next disease d_i is estimated via the formula

$$\Pr\left(t_{w+1}^i\right) = \underset{d_i \in D_{C_i}}{\operatorname{argmax}} \left\{ \Pr\left(t_{w+1}^i = d_i | t_w^i, t_{w-1}^i, \ldots, t_1^i\right) \right\} \tag{8}$$

In practice, d_i is accepted as the most probable next disease only if it results in a state whose probability is significantly better than that of the second most probable predicted disease.

More in details, the $100(1 - \alpha)$ percent confidence interval around the most probable next disease is computed, and thus it is checked if the second predicted disease falls within this interval [14]. If this condition happens, the most probable state is discarded, otherwise it is accepted as next predicted disease. More in detail,

Table 6 2nd-order transition probability matrix corresponding to the patient medical records of Table 4

2nd order	401	437	715	722	723	756
{401,715}	0	0	0	1	0	0
{401,437}	0	0	2	0	0	0
{437,715}	0	0	0	3	0	0
{715,722}	0	0	0	0	2	2
{722,723}	0	0	0	0	0	0
{722,756}	0	0	0	0	0	0

Table 7 Set T of patient records involving some common diseases

t_1	{401, 715, 722, 723}
t_2	{401, 437, 715, 722, 756}
t_3	{437, 592, 715, 722, 723}
t_4	{401, 437, 592, 715, 722, 723}
t_5	{437, 715, 722, 756}
t_6	{592, 715, 722, 723, 756}
t_7	{401, 592, 715, 722, 756}
t_8	{401, 437, 715, 721, 722, 723}
t_9	{241, 255, 595, 780}
t_{10}	{241, 255, 272, 595, 780}

if $\hat{p} = \Pr(d_i)$ is the probability of the most probable disease, then its $100(1 - \alpha)$ percent confidence interval is given by

$$\hat{p} - z_{\alpha/2} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{9}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution and n is the frequency of the Markov state [15].

In the case the Markov model is unable to provide us with a reliable prediction, association rules are used to circumvent the problem. In particular, the patient medical record t_i^w is matched against all the frequent itemsets $I_{C_i}^{w+1}$ of size $w + 1$ (i.e., all those itemsets whose support *supp* is greater than a fixed threshold σ) induced on C_i. Each itemset $i_i^{w+1} \in I_{C_i}^{w+1}$ containing t_i^w contributes to the set of the candidate diseases with a prediction d_i. It is easily noticing that $i_i^{w+1} = t_i^w \cup \{d_i\}$. Finally, if the confidence of the rule $t_i^w \Rightarrow \{d_i\}$ (i.e., $supp(t_i^w \cup \{d_i\})/supp(t_i^w)$) is greater than a fixed threshold τ, the disease d_i is considered reliable, and it is added to the set of predicted diseases.

5 A running example

In order to explain the way our prediction approach works in practice, let us consider the set T of patient medical records reported in Table 7.

Let us also suppose $k = 2$ be the number of clusters the clustering algorithm is forced to find, and $\gamma = 0.5$ be the minimum percentage of occurrences a disease must have for being inserted into the cluster representative (see Eq. 3). On the base of the above parameters, it is easily verifiable that the clustering algorithm finds the clusters C_1 and C_2, as reported in Tables 8 and 9, respectively. Furthermore, the clusters are equipped with their representatives: $r_{C_1} = \{401, 437, 592, 715, 722, 723, 756\}$ and $r_{C_2} = \{241, 255, 272, 595, 780\}$. After the clusters have been built, the disease prediction model is carried out for each cluster found.

Now, let $t = \{t_1 = 715, t_2 = 722\}$ be a new patient disease record. Since the distance $d(t, r_{C_i}) = 1 - 2/7 = 0.714$ is lower than $d(t, r_{C_2}) = 1, t$ is recognized belonging to C_1, thus the Markov model built upon C_1 is exploited to perform the

Table 8 Cluster C_1

t_1	{401, 715, 722, 723}
t_2	{401, 437, 715, 722, 756}
t_3	{437, 592, 715, 722, 723}
t_4	{401, 437, 592, 715, 722, 723}
t_5	{437, 715, 722, 756}
t_6	{592, 715, 722, 723, 756}
t_7	{401, 592, 715, 722, 756}
t_8	{401, 437, 715, 721, 722, 723}

Table 9 Cluster C_2

t_9	{241, 255, 595, 780}
t_{10}	{241, 255, 272, 595, 780}

predictions. In particular, if the window size w is set to 2, the 2nd-order Markov model is picked to be used.

It can be easily noticing that in Table 8 the state {715, 722} appears 7 times, while diseases 723 and 756 appear 4 times and 3 times after this state, respectively. Thus:

$$\Pr(t_3) = \text{argmax}\{\Pr(t_3 = 723 | t_2 = 722, t_1 = 715)\} = \text{argmax}\{t_3 = 723 \rightarrow 0.57\}$$

and

$$\Pr(t_3) = \text{argmax}\{\Pr(t_3 = 756 | t_2 = 722, t_1 = 715)\} = \text{argmax}\{t_3 = 756 \rightarrow 0.43\}$$

However, this information does not necessarily provide us with the the correct prediction of next disease since there is not an enough probability difference for the diseases 723 and 756 (see Sect. 4). More in detail, if we compute the confidence interval for the most probable next disease at 90 % confidence level (i.e., $z_{\alpha/2} = 1.65$), we obtain that this may vary approximately between 0.33 and 0.85. Note that, because of the small number of instances in this example, the greater the confidence level, the larger the confidence interval. Since the probability of the other disease $\Pr(t_3 = 756) = 0.43$ falls in this interval, we cannot consider the prediction made by the Markov model reliable. In this case of uncertainty, in order to disambiguate the choice, we resort to the predictive capability of the association rules. If we fix the threshold support for the itemsets to 0.5, the frequent itemsets for C_1 are all those having a support $\sigma > 4$ (see Table 10).

By matching the medical record $t = \{715, 722\}$ against the 3-frequent itemsets I^3, both the diseases having codes 723 and 756 are candidate for being the likely, next diseases the patient t may incur in. However, the diseases 723 and 756 change their status from candidate to predicted only if the confidence of the association rules ar_1: $\{715, 722\} \Rightarrow \{723\}$ and ar_2: $\{715, 722\} \Rightarrow \{756\}$ exceeds the minimum confidence threshold. Therefore, if we set the threshold for the confidence to 0.6, since the confidence of ar_1 is

$$\tau(ar_1) = \sigma(\{715, 722, 723\})/\sigma(\{715, 722\}) = 5/8 = 0.625$$

and the the confidence of ar_2 is

Table 10 Frequent itemsets built upon cluster C_1

I^1	I^2	I^3
{401} (5)	{401, 715} (5)	{401, 715, 722} (5)
{437} (5)	{401, 722} (5)	{437, 715, 722} (5)
{592} (4)	{592, 715} (4)	{592, 715, 722} (4)
{715} (8)	{437, 715} (5)	{715, 722, 723} (5)
{722} (8)	{437, 722} (5)	{715, 722, 756} (4)
{723} (5)	{715, 722} (8)	
{756} (4)	{715, 723} (5)	
	{715, 756} (4)	
	{722, 723} (5)	
	{722, 756} (4)	
	{592, 722} (4)	

$$\tau(ar_2) = \sigma(\{715, 722, 756\})/\sigma(\{715, 722\}) = 4/8 = 0.5$$

only the disease 723 is definitively added to the set of predicted illnesses. Therefore, by means of the rule ar_1, we foresee that a patient having osteoarthrosis (715), and disc disorders (722), is very likely to present the symptom of other disorders of cervical region (723) in the future.

In the next section we will show at what extent *CORE* is effective in predicting, next, incoming diseases.

6 Data description

The data set consists of the medical records of two small cities in the south of Italy. Each record contains a unique patient identifiers, its birth date, the gender, and the list of disease codes with the date of the visit in which that disease has been diagnosed. The disease codes are those defined by the *International Classification of Diseases, Ninth Revision, Clinical Modification* (*ICD-9-CM*). The *International Classification of Diseases (ICD) and Related Health Problems* supplies codes to classify diseases and a wide variety of signs. The ICD is published by the *World Health Organization* and used worldwide for morbidity and mortality statistics, reimbursement systems and automated decision support in medicine. Every health condition is associated with a unique category and given a code, up to five digits long. The first three digits constitute the principal diagnosis, while the other two identify secondary diagnoses.

For our purposes, we kept only the first three digits of each ICD-9-CM code since they are informative enough to analyze the disease correlations we are interested in. Patients having no one or only one diagnosis have been been discarded because not useful at all. After this preprocessing step, tge total number of patients is 2,541 (1,105 and 1,436 coming from the first and second town, respectively), and the number of distinct diseases is 455.

Table 11 The top 20 most recurrent diseases

	Disease	Prevalence (%)
1	401 *(Hypertension)*	33.16
2	530 *(Diseases of esophagus)*	22.02
3	715 *(Osteoarthrosis and allied disorders)*	21.54
4	722 *(Intervertebral disc disorders)*	17.72
5	462 *(Pharyngitis)*	14.81
6	250 *(Diabetes mellitus)*	14.26
7	466 *(Acute bronchitis and bronchiolitis)*	11.70
8	733 *(Other disorders of bone and cartilage)*	10.32
9	724 *(Other and unspecified disorders of back)*	8.90
10	464 *(Acute laryngitis and tracheitis)*	8.86
11	721 *(Spondylosis and allied disorders)*	8.07
12	240 *(Simple and unspecified goiter)*	8.07
13	272 *(Disorders of lipoid metabolism)*	6.58
14	595 *(Cystitis)*	6.54
15	535 *(Gastritis and duodenitis)*	6.42
16	427 *(Cardiac dysrhythmias)*	6.18
17	600 *(Hyperplasia of prostate)*	6.07
18	300 *(Anxiety, dissociative and somatoform disorders)*	5.04
19	491 *(Chronic bronchitis)*	4.84
20	726 *(Peripheral enthesopathies and allied syndromes)*	4.81

Table 11 shows the 20 most recurring diseases in our data set. It is interesting to note that, though our data set has a different geographical origin, and it is rather small with respect to that studied by Davis et al. [6, 7]—which contains 13 millions of patients—the prevalence of some diseases are almost the same. For instance, *Hypertension, Diabetes mellitus, Cardiac dysrhythmias*, and *Hyperplasia of prostate* have a prevalence in percentage of 33.16, 14.26, 6.18, and 6.07 % in our data set, while, in the data set used in [7], the prevalences are 33.64, 10.47, 5.61, and 6.54 %, respectively. Please, notice that the latter data set is not publicly available.

7 Evaluation measures

In this section we describe the evaluation measures used to test the effectiveness of our approach, and the methodology employed to compute such values.

In order to perform a fair evaluation, we applied the well-known 10-fold *cross validation* method [5], i.e., the original dataset is split in 10 equal-sized partitions. During each of the 10 runs, one of the partitions is chosen for testing, while the rest of them are used for training the prediction model. The cumulative error is found by summing up the errors for all the 10 runs. The strategy we followed for testing our approach is detailed in the following.

First of all, the records in the training set T_{train} are partitioned in k clusters, and for each group, a distinct prediction model M_i is built upon. Relatively to the dataset at hand, we empirically found that $k = 10$ (number of clusters) and $\gamma = 0.5$ (threshold value for the representative computation) is the setting that ensures the best partitioning for the dataset at hand, i.e., that minimizing the cost function Q_k (see Eq. 1). A record t in the test set T_{test} is assigned to one of the k cluster, then it is divided in two subsets of diseases, $head_t$ and $tail_t$.

The subset of diseases $head_t$ is used for generating predictions, while the remaining diseases in $tail_t$ are used to evaluate the prediction. It is worth noticing that the length of $head_t$ is tightly related to the maximum window size w allowable for each cluster, and, intuitively, it must be lower than the maximal length of frequent itemsets mined in each cluster. In our dataset, since the frequent patterns generated are of size at most 5, the maximum length of $head_t$ cannot exceed 4.

More in general, given a record t and a window size w, we select the first w diseases of t as $head_t$ and the remaining $|t| - w$ as $tail_t$. If the record t belongs to the cluster C_i, the relative prediction model M_i first applies the wth-order Markov model $MM^w_{C_i}$. If either $head_t$ does not match any state of $MM^w_{C_i}$ or its prediction has a low probability (see Sect. 4), association rules for predicting the next diseases are exploited instead. To this purpose, $head_t$ is matched against all patterns $I^{w+1}_{C_i}$ (i.e., the frequent itemsets of length $w + 1$ induced in the cluster C_i). Note that, as previously explained, low probability means that the second predicted disease falls within the confidence interval of the first predicted disease [14]. To compute the confidence interval we used a value of $z_{\alpha/2} = 1.65$. This value has been obtained by properly tuning $z_{\alpha/2}$ on the data set at hand. In fact, since the greater the confidence level, the larger the confidence interval, if we use high confidence levels it is more likely that the second predicted disease will fall in the $100(1 - \alpha)$ confidence interval, thus we would need to recur to association rules more often than to Markov models for performing the next prediction. As a direct consequence of this behavior, the advantage deriving by the adoption of Markov models as first stage of our *prediction* step would be completely lost. For this reason, we decided to adopt a less strict constraint for the confidence interval because we empirically found that this value avoids to exclude the Markov models form the prediction phase, so allowing a seamless cooperation of both predictors in order to achieve the best accuracy possible.

Let P_{head_t} be the set containing all the candidate predictions made by either exploiting Markov models or association rules. In this latter case, fixed the minimum confidence threshold τ, P_{head_t} will contain all the candidate predictions whose confidence is greater than τ. Subsequently, the set P_{head_t} is compared with $tail_t$ in order to validate the prediction. The comparison of these sets is done by using two different metrics, namely *Precision* and *Recall* [5]. Precision and recall are two widely used statistical measures in the data mining field. In particular, precision is seen as a measure of exactness, whereas recall is a measure of completeness.

By customizing these definitions to our scenario, we exploited precision for assessing how accurate the provided predictions are (i.e., the proportion of relevant

predictions to the total number of predictions) and recall for testing if we predicted all the diseases the patients are likely to be affected in the future (i.e, the proportion of relevant predictions to all diseases that should be predicted). Formally, the precision of P_{head_t} is defined as

$$Precision(P_{head_t}) = \frac{|P_{head_t} \cap tail_t|}{|P_{head_t}|} \tag{10}$$

and the recall of P_{head_t} as

$$Recall(P_{head_t}) = \frac{|P_{head_t} \cap tail_t|}{|tail_t|} \tag{11}$$

Another metric, the *F-measure* [5], which is the harmonic mean between precision and recall, is often used to examine the tradeoff between precision and recall:

$$F-\text{measure} = \frac{2 * \text{precision} * \text{recall}}{\text{recall} + \text{precision}} \tag{12}$$

In the next section the values of precision, recall, and f-measure will be computed to estimate the prediction capability of our system.

8 Results

In this section we present the results and evaluate them on the base of the introduced metrics. Figure 2a, b show the cumulative precision and recall scores obtained by fixing the overall support σ for the frequent patterns to 0.1. This choice has been motivated by the fact that, since the size of the data set is not so large, a low support value is necessary for ensuring an adequate length for the mined patterns, also in the case of poor cluster homogeneity. The values of both precision and recall have been computed by varying the confidence threshold τ between 0.1 and 1, and the window size w from 2 to 4. Increasing window size allows for evaluating the impact of the number of considered diseases on the quality of the results. Please, notice that we do not consider the too restrictive case of $w = 1$, i.e., just one disease in the patient's medical history used for the prediction.

Figure 2a clearly reveals that precision increases as a larger portions of patient medical history, i.e. an increasing number of diseases, are used to compute predictions, reaching 0.81 when a window of size 4 and a confidence value 1 are used. Conversely, recall is negatively biased by larger window sizes, as pointed out in Fig. 2b, though the recall reduction rate from size 4 to size 2 is low. For example, when the confidence is 1, it diminishes from 0.49 to 0.45.

Figure 3 displays instead the behavior of precision and recall when the support threshold σ varies, and the window size w has been fixed to 4.

Increasing the support threshold has two main positive effects: *(i)* improving the precision of predictions, and *(ii)* ensuring the scalability of the association rule mining algorithm, since a lower number of frequent itemsets is computed. However, as a side effect, a higher support results in a potential loss of some important, yet

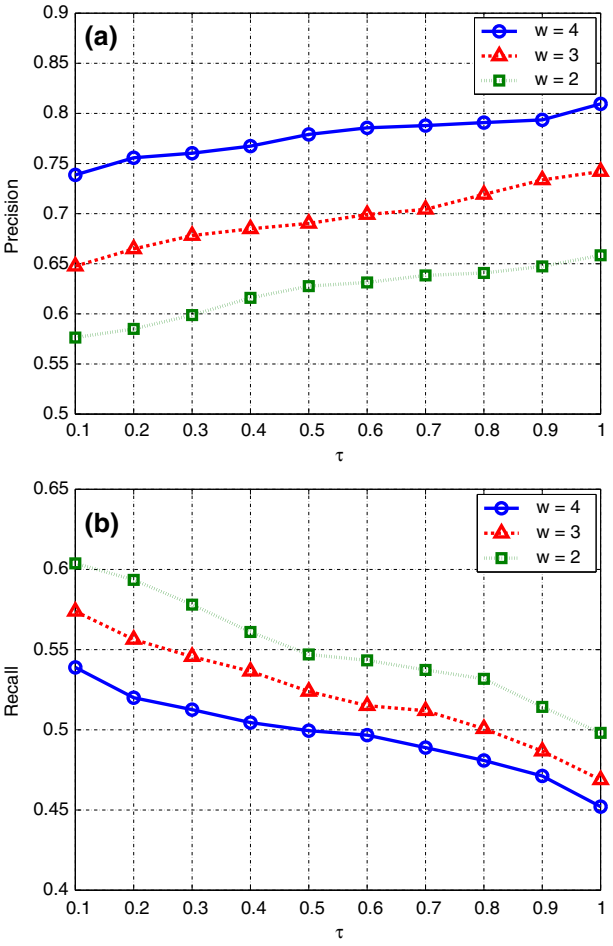

Fig. 2 Impact of w on precision and recall measures when $\sigma = 0.1$ and confidence $\tau \in [0.1, 1]$

infrequent, diseases in the prediction set. In the medical context, this kind of illnesses could be particularly important and more informative for producing a correct diagnosis. Figure 3a clearly points out better performances of precision when the support threshold increases. Indeed, it is easy to note that, for $\tau = 1$, we obtain a precision of 0.81 if $\sigma = 0.1$, 0.82 if $\sigma = 0.15$, and 0.84 for $\sigma = 0.2$. An inverse trend can be noted in Fig. 3b for the recall which progressively decreases from 0.54 to 0.45, when $\sigma = 0.1$, from 0.50 to 0.44, when $\sigma = 0.15$, and from 0.48 to 0.42, when $\sigma = 0.2$, respectively.

For the sake of completeness, Fig. 4 shows the *F-measure* values when w varies from 2 to 4, $\sigma = 0.1$, and τ ranges in [0.1, 1]. The figure points out that with a window size of 2 and a confidence value <0.4 gives a F-measure value higher than that obtained with a larger window size and confidence equal to 1. The choice of

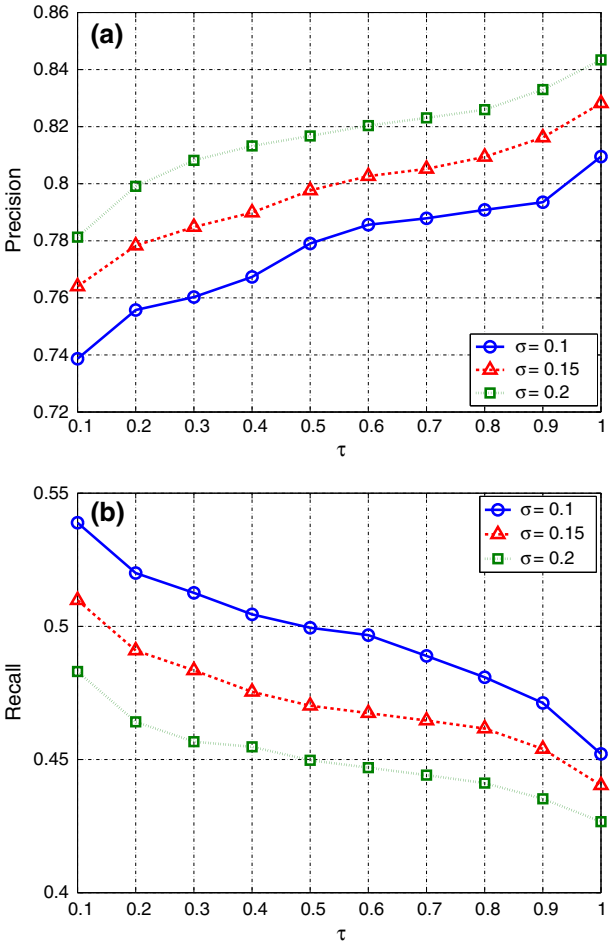

Fig. 3 Impact of σ on precision and recall measures when $w = 4$ and confidence $\tau \in [0.1, 1]$

parameter values can thus be done by evaluating the trade-off between enlarging the window size w and reducing confidence value τ.

Finally, as a practical result, we show some of the most relevant predictions the *CORE* system is able to foresee. In particular, Table 12 reports some of the association rules found by the system, ordered with respect to their confidence value. Thus, for example, the first rule states that if a patient is affected by *Asphyxia*, *Acute bronchitis and bronchiolitis*, he could get also *Asthma* in the next future. A confidence value equal to 1 means that, relatively to the data set of medical records used, the disease contained in the right part of the rule always appears together with the diseases of the left part. Note that, since *Hypertension* is the most prevalent disease in the medical records of our experimentation, this illness appeared many times in the head of the computed association rules—Table 12 just reports two of these rules.

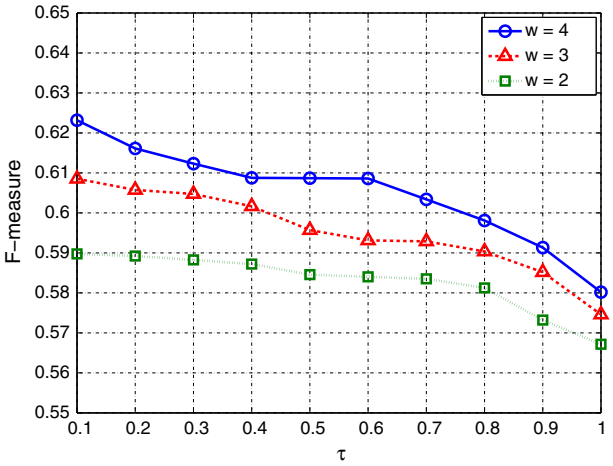

Fig. 4 *F*-measure when $w = 2, 3, 4, \tau \in [0.1, 1], \sigma = 0.1$

Table 12 Association rules returned by the *CORE* system by setting $w = 4$, $\sigma = 0.1$, and $\tau = 0.4$

Association rule	Confidence
Asphyxia, Acute bronchitis and bronchiolitis → Asthma	1
Other disorders of cervical region, Curvature of spine → Osteoarthrosis and allied disorders	1
Other disorders of cervical region, Spondylosis and allied disorders → Intervertebral disc disorders	1
Other disorders of arteries and arterioles → Essential Hypertension	0.87
Other of back, Osteoarthrosis and all. dis., Spond. and all. dis. → Other dis. of bone and cart	0.83
Intervertebral disc disorders, Abnormal red blood cell → Other disorders of bone and cartilage	0.75
Asthma, Asphyxia, Pharyngitis acute → Anaphylaxis	0.75
Atherosclerosis, Cardiac dysrhythmias → Essential Hypertension	0.72

The *CORE* system, thus, besides predicting the diseases of a patient, given his historical medical record, can also provide the physician with a set of rules explaining the performed predictions.

9 Conclusions

A method based on the combination of *clustering, Markov models*, and *association analysis* for the prediction of incoming diseases has been presented. Basically, the approach uses the past medical history of patients to determine next diseases an individual could incur in the future.

As the experimental results proved, the combination of more models allows for a good prediction accuracy. The technique can thus be considered as a feasible approach to the disease prediction. It is worth noting that the main limitations to our analysis come from the inherent characteristics of data set at hand, i.e., (a) small size, (b) locality, i.e., the strict provenance of patients from a small area in the south of Italy, and (c) patient age. In fact, we did not restrict our analysis to particular groups of people—e.g., elderly patients as other approaches do [6, 7]—but we considered the whole data set. All these factors could have biased the results, for example because of the probable presence of health problems specific for the geographical area considered, or due to the patient age. Furthermore, it is worth noting that the access to patient medical records is very difficult, mainly due to privacy problems and to the reluctance of physicians and medicare systems to make their data publicly available. Despite these weaknesses, experimental results showed that the combination of knowledge discovery techniques is a promising way to advance the disease prediction.

Currently, the Italian care system is making an important effort towards a profitable usage of electronic patient records, thus, as a future work, we will try to extend the study to a much larger population, possibly exploiting more specific patient information—beside the diagnosed diseases—for improving the prediction quality.

References

1. Hidalgo CA, Blumm N, Barabási A-L, Christakis NA (2009) A dynamic network approach for the study of human phenotypes. PLoS Comput Biol 5(4):e1000353. doi:10.1371/journal.pcbi.1000353
2. Starfield B et al (2003) Comorbidity: implications for the importance of primary care in 'case' managment. Ann Family Med 1(1):8–14
3. Lowensteyn I et al (1998) Can computerized risk profiles help patients improve their coronary risk? The results of the coronary health assessment study. Prev Med 27(5):730–737
4. Wilson PWF et al (1998) Prediction of coronary heart disease using risk factor categories. Circulation 97:1837–1847
5. Tan P, Steinbach M, Kumar V (2006) Introduction to data mining. Pearson International Edition, San Francisco
6. Davis DA, Chawla NV, Blumm N, Christakis NA, Barabási A-L (2008) Predicting individual disease risk based on medical history. In: Proceedings of the ACM international conference on information and knowledge management (CIKM'08), pp 769–778
7. Davis DA, Chawla NV, Christakis NA, Barabási AL (2010) Time to CARE: a collaborative engine for practical disease prediction. Data Mining Knowl Discov 20:388–415
8. Shardanand U, Maes P (1995) Social information filtering: algorithms for automating word of mouth. In: Proceedings of ACM conference on human factors in computing systems (CHI'95), pp 210–217
9. Steinhaeuser K, Chawla NV (2009) A network-based approach to understanding and predicting diseases. In: Social computing and behavioral modeling
10. Strehk RMA, Ghosh J (2000) Impact of similarity measures on web-page clustering. In: Proceedings of AAAI workshop on AI for web search, pp 58–64
11. Jaccard P (1912) The distribution of the flora of the alpine zone. New Phytol 11:37–50
12. Giannotti F, Gozzi C, Manco G (2002) Clustering transactional data. In: Proceedings of principles of data mining and knowledge discovery (PKDD'02), pp 175–187
13. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th berkeley symposium, vol 1, pp 281–297
14. Deshpande M, Karypis G (2004) Selective Markov models for predicting web page accesses. ACM Trans Internet Techn 4(2):163–184
15. Mongomery D, Runger G (2004) Applied statistics and probability for engineers. Wiley, London