基于示例的组合预测方法*

巩昌平 陆玉昌 周远晖

(清华大学计算机系 北京 100084)

摘要 本文提出了一种基于示例的组合预测方法,强调知识方法和数学方法的结合,提出了一种算法和组合预测框架,并结合实验数据讨论了预测结果,分析了不同预测方法的不足.

关键词 示例,分类,匹配,组合预测.

以概率、统计为基础的因果回归分析法和时间序列分析方法被广泛应用于预测,其前提条件是发展历史蕴涵未来,从而通过对历史数据的处理可以估计未来问题.但这些方法对国家政策、地方政策、地方保护、假货进入等突发事件不敏感,起码是滞后一段时间;在回归分析中,存在许多未知的因素,且这些因素是随时间变化的.因而给预测的正确性带来许多困难,也就影响了管理决策.

分析发现,各影响因素对各地区而言常是不同时起作用的,地方法规的制定更是有先有后,对假货进入情况也存在地区差和时间差.这就启发我们,通过类比、匹配、相似性判别等手段,找出这种地区差和时间差,用建立典型示例的手法进行预测,这就是基于示例的预测法,其理论基础是基于示例的推理 CBR(case-based reasoning).

基于示例的推理 CBR 是美国 Yale 大学 Roger C. Sckank 教授在 1982 年《Dynamic Memory》中提出的,基本思想是从记忆的典型示例中选择与新示例相似的、进而借助于典型示例的有关信息来对新示例进行解释.CBR 对处理领域知识不丰富的问题,有独到的优势.直观地说,基于示例的预测有如"天津"的明天,可由"北京"的今天来描述.

为了收到预测的较好效果,需对预测对象提出几种不同的预测方案,在各种方案中,充分衡量预测对象变化的条件以及可能变化的幅度,并采取相应的处理方法.基于此,本文给出了一种知识方法与数学方法相结合的组合预测方法 CBCF(case-based combinatorial forecast).

1 基于示例的组合预测模型

定义 1.1. 对一连续曲线 $x(t)$来说,分成 n 段,分点 t_i 的函数值为 $x(t_i)$,简记为 x_i,则 $(x_1,\cdots,x_n)$称为示例集.

---

* 作者巩昌平,1964 年生,博士生,主要研究领域为定性推理.陆玉昌,1937 年生,教授,主要研究领域为人工智能.周远晖,1969 年生,博士生,主要研究领域为人工智能与应用.

本文通讯联系人:巩昌平,北京 100084,清华大学计算机系

本文 1996-01-30 收到修改稿

定义 1.2. 在示例集中任取一段序列，且保持示例集中的序关系，称为该示例集中的示例. 示例集中角标最大的一段所构成的示例称为最新示例，否则，称为一般示例，示例与示例之间允许相交. 例如，(x_1,x_2,x_3)，(x_{n-2},x_{n-1},x_n) $(n\geqslant 3)$ 为示例集 $(x_1,x_2,\cdots,x_n)$ 中的示例，其中 (x_{n-2},x_{n-1},x_n) 为最新示例，(x_1,x_2,x_3) 为一般示例.

定义 1.3. 一组示例集组成一个集类，称为总体示例集.

定义 1.4. 同一示例集中，示例的相似称为自相似；不同示例集中，示例的相似称为它相似.

以预测对象的历史数据，按时间序作为示例集，该示例集的波动规律代表了发展变化. 因而示例集中任一段子序列可以看成一个样本，构成示例集中的示例，示例集中最新的一段所构成的示例称为最新示例. 例如，北京地区烟草销售量按月的数据构成一示例集 $(x_{t-n+1},\cdots,x_{t-2},x_{t-1},x_t)$，其中 t 表示当前时间，x_{t-i} $(i=1,\ldots,n-1)$ 表示前 i 月的数据，其中 (x_{t-2},x_{t-1},x_t) 为最新示例，$(x_{t-4},x_{t-3},x_{t-2})$ 为一般示例. 不同地区，例如北京、天津、上海……构成不同的实例集，将它们合在一起构成总体示例集，即 $\{(x^1_{t-n+1},\cdots,x^1_{t-2},x^1_{t-1},x^1_t),(x^2_{t-n+1},\cdots,x^2_{t-2},x^2_{t-1},x^2_t),\cdots\}$，其中 x^j_t 表示 t 时的 j 地区.

当要预测某一地区的情况时，如“天津”，先选取该地区的最新示例，然后在总体示例集中，根据一定的策略选取与该最新示例最“相似的”，例如，经过一定策略的判定，天津地区的最新示例 $(x^2_{t-2},x^2_{t-1},x^2_t)$ 与北京地区的示例 $(x^1_{t-4},x^1_{t-3},x^1_{t-2})$ 相似，因而可以用北京地区的示例 $(x^1_{t-4},x^1_{t-3},x^1_{t-2})$ 的延伸 (x^1_{t-1},x^1_t)，作为 (x^2_{t+1},x^2_{t+2}) 的估计.

基于示例的预测结果也有一定的局限性，如可能没有好的相似示例；预测的结果从理论上不能保证收敛到实际结果等. 而数学方法与知识方法相结合，进行组合预测(CBCF)更为有效.

CBCF 是把基于示例的预测与数学方法、知识方法相结合进行的预测；是对不同的预测结果进行加权，根据专家经验知识，对各预测方法作用的大小做一评价，确定一权函数，通过调整权函数来给出预测结果. 实验表明效果较好，弥补了一些非可知因素对预测结果的影响.

2 预测方法框图、算法流程

基于示例的推理 CBR 的运行机理，可表示为图 1.

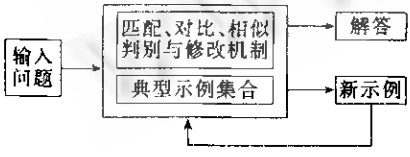

图1

根据输入问题的特征，CBR 系统的索引机制在典型示例库中搜寻相似的候选示例集合，由与问题的匹配程度决定可用的示例，然后运用相似示例的有关信息指导问题的求解. 以玉溪卷烟厂信息决策支持系统中预测为例说明其算法.

已知总体示例集 $C=\{(x^1_{t-n+1},\cdots,x^1_{t-2},x^1_{t-1},x^1_t),(x^2_{t-n+1},\cdots,x^2_{t-2},x^2_{t-1},x^2_t),\cdots\}$，其中每一元素代表某一地区预测对象的历史数据构成的示例集合，示例 *case* 为任一地区的某一段子序列，新示例 *newcase* 为任一地区的最后一段子序列，示例的长度相等.

CBR **算法**

1. 对总体示例集根据某原则进行分类,属于同一类的才能进行相似分析(匹配),分类后的总体示例集记为(C,S),S 表示分类结构.

2. 若(C,S)非空,选一类(C,S_0);若(C,S)为空,结束.

3. $(C,S)-(C,S_0)$,"−"表示删除;同一类中,若存在新示例 *newcase*,选一个;若不存在,goto 2.

4. $(C,S_0)-newcase$;进行相似判断,即在(C,S_0)中选一候选 *case* 通过匹配、相似判断,找一合适的;若找不到合适的,找最好的.注意:候选 *case* 和 *newcase* 若在同一示例集中,表示自相似 *case* 都不是最新示例.

5. 匹配上的 *case* 的后续,作为 *newcase* 的预测结果.

6. goto 3.

以上算法结束,各地区也遍历了一遍,除非在分类中单独一类,则每个地区都可获得预测值.若某地区单独一类,用 ARMA 模型或其它平滑方法单独处理.

CBCF 算法

1. 建立 ARMA 模型、回归模型.

2. 利用所建数学模型进行各地区的预测.

3. CBR 预测算法.

4. 通过知识、专家经验公式建立权函数,进而建立组合预测模型.

5. 利用组合预测模型进行组合预测.

CBCF 算法的数据流图由图 2 描述,实际程序的模块调用关系如图 3.

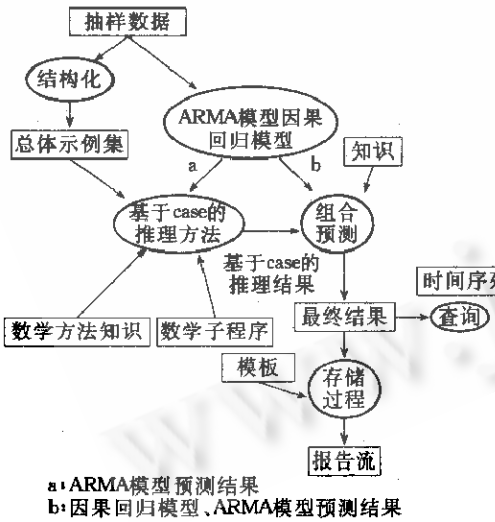

图2

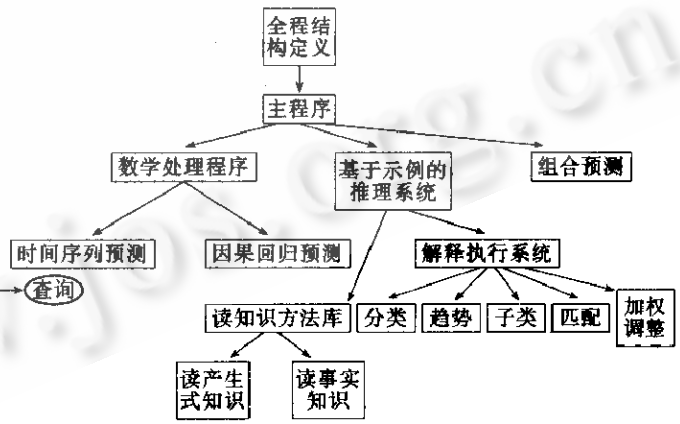

图3

图 2、图 3 描述了预测系统的基本骨架、通过添加不同的预测对象和知识能产生丰富的预测结果.

3　分类、匹配

我们考察了烟草行业在不同地区消费特点的差异,例如,沿海和靠香港的经济发达地区

外烟销量明显较高,而在内地却有不同.由于地方保护主义的影响,有些地区政府要求以消费本地生产的烟为主,另有一些地区喜欢吸玉溪卷烟厂的烟等.根据这一消费特点的差异,我们将总体示例集划分成 4 类,第 1 类是以消费玉溪烟为主的地区;第 2 类以消费本地烟为主的地区;第 3 类为以消费外烟为主的地区;第 4 类为以消费构成玉溪威胁的竞争厂家烟为主的地区我们规定不同消费特点的地区不作相似匹配.在这 4 类的基础上,通过趋势分析,再划分子类,例如趋势走向为上升的为一子类,趋势走向为下降的为另一子类,从而得到 8 个子类.这样,一方面匹配结果反映了消费特点,另一方面是给总体示例集赋予了一种结构,在对候选示例识别时,对所得结果可做解释,更有意义.

匹配指的是典型示例 *case* 和新示例 *newcase* 的相似程度,可从 2 方面处理匹配问题,一是以概率统计为基础的相关分析,这时发展趋势接近,波动情况相似的示例得以匹配,这在一定程度上反映了该地区的消费心理和季节变化.另一方面是通过 2 地区基本经济情况、消费心理、对国家政策执行的先后顺序、地方保护的变化进行匹配,这时易于用知识方法描述.总的原则是某一地区的过去,确实与另一地区的今天有一定的相似性.

4 组合预测

CBR 方法不一定能找到好的匹配,而且 CBR 得到的预测结果不能保证从理论上收敛到实际结果.CBCF 算法强调 CBR 和数学方法、专家经验知识的结合,即组合预测.组合预测的模型,主要是加权算法,权函数的确定是至关重要的.

权函数的调整由专家系统实现,根据各地区的不同特点,运用专家的知识和经验公式,对权函数进行调整.同时,提供一个交互式、可验证预测结果的环境,比如,留几个最新数据不参与预测,和预测的结果相比较,用户根据比较结果反馈到预测模型中去,以提高预测的质量.这种方法是行之有效的,因为用历史数据作预测,存在着固有的缺点.如,若国家政策发生重大变化,厂方的生产、经营计划大幅度改变,用历史数据预测是无法得到反映的,因而预测的结果会滞后很长一段时间,才能反映出来.而若把这些作为知识,启动专家系统作综合评判,把能反映这一变化的预测方法加大权重,提高当前历史数据的价值,肯定会得出更令人满意的预测结果.而这些信息,用户是容易得到的,因而可以达到事半功倍.

5 实验结果

通过对 94 年烟草业 1～9 月销售数据的预测结果看,CBCF 方法比单一方面有明显的改进,结果也令人满意.用 1～7 月数据作预测,用 8、9 月的数据作检验,结果大部分预测结果达到误差小于 5%.对数据缺乏的地区,经插补后预测结果也明显得到改善.图 4 为安徽省的预测结果,纵坐标表示销售量,横坐标表示时间.图 4 的上面部分为安徽省,下面部分为匹配示例集湖北省,匹配结果为湖北省超前安徽省 2 个月,因而可以把湖北省 6、7 月份的数据作为安徽省 8、9 月份的基于示例的预测数据.当前时间右边为组合预测数据以条带的形式给出.

致谢 本文在石纯一教授的直接指导下完成,北京大学陈良猛教授,中国人民大学张尧庭教

授给予了具体的建议，在此深表谢意.

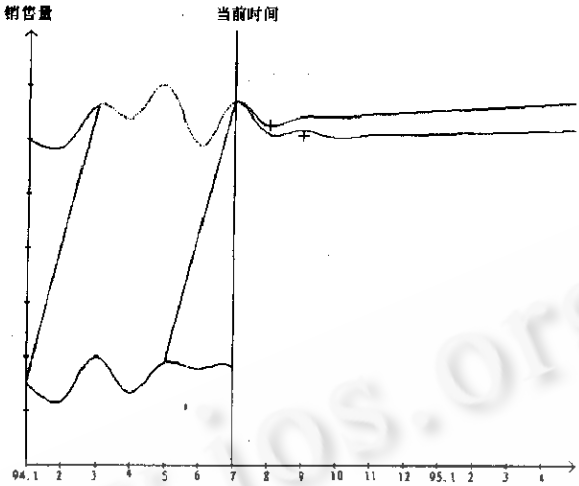

图 4

参考文献

1　李子奈. 计量经济学. 北京：清华大学出版社，1992.
2　顾岚. 时间序列在经济中的应用. 北京：中国统计出版社，1993.
3　易丹辉. 统计预测. 北京：中国人民大学出版社，1988.
4　Yves Meger. 小波与算子. 世界图书出版公司，1992.
5　Hammond K J CHEF. A model of case—based planning. AAAI，1986.
6　Ram A. Indexing elaboration and refinement：incremental learning of explanatory case. Machine learning，1993，10：201～248.

CASE—BASED COMBINATORIAL FORECAST

Gong Changping　Lu Yuchang　Zhou Yuanhui

(*Department of Computer Science　Tsinghua University　Beijing*　100084)

Abstract　This paper presents a general framework and algorithms for a case—based combinatorial forecast method，which emphasizes the combining of knowledge and mathematics. By analyzing expermental results，the paper discusses the shortcomings of various methods of forecast.

Key words　Case，class，match，combinatorial forecast.